Special issue: comparative survey analysis – models, techniques, and applications

Meuleman, Bart ; Davidov, Eldad ; Seddig, Daniel

mda

methods, data, analyses

Journal for Quantitative Methods and Survey Methodology

Volume 12, 2018 | 2

Comparative Survey Analysis: Models, Techniques, and Applications

Bart Meuleman, Eldad Davidov, & Daniel Seddig (Editors)

Bart Meuleman et al.	Modeling Multiple-country Repeated Cross-sections
Christoph Spörlein & Elmar Schlueter	Demonstrating How to Best Examine Group-based Segregation
Dominik Becker et al.	Surpassing Simple Aggregation
Nate Breznau	Simultaneous Feedback Models with Macro-Comparative Cross-Sectional Data
Christian Schnaudt & Michael Weinhardt	Blaming the Young Misses the Point

methods, data, analyses is published by GESIS – Leibniz Institute for the Social Sciences.

Managing editor: Sabine Häder
GESIS – Leibniz Institute for the Social Sciences
PO Box 12 21 55
68072 Mannheim
Germany
Tel.: + 49.621.1246282
E-mail: mda@gesis.org
Internet: www.mda.gesis.org

Methods, data, analyses (mda) publishes research on all questions important to quantitative methods, with a special emphasis on survey methodology. In spite of this focus we welcome contributions on other methodological aspects. We especially invite authors to submit articles extending the profession's knowledge on the science of surveys, be it on data collection, measurement, or data analysis and statistics. We also welcome applied papers that deal with the use of quantitative methods in practice, with teaching quantitative methods, or that present the use of a particular state-of-the-art method using an example for illustration.

All papers submitted to mda will first be screened by the editors for general suitability and then double-blindly reviewed by at least two reviewers. Mda appears in two regular issues per year (January and July).

Print: Bonifatius Druck GmbH Paderborn, Germany

ISSN 1864-6956 (Print)
ISSN 2190-4936 (Online)

Content

181 Editorial:
Comparative Survey Analysis – Models, Techniques and Applications
Bart Meuleman, Eldad Davidov, Daniel Seddig

RESEARCH REPORTS

185 Modeling Multiple-country Repeated Cross-sections. A Societal Growth Curve Model for Studying the Effect of the Economic Crisis on Perceived Ethnic Threat
Bart Meuleman, Eldad Davidov, Jaak Billiet

211 Demonstrating How to Best Examine Group-based Segregation: A Statistical and Conceptual Multilevel Approach
Christoph Spörlein, Elmar Schlueter

233 Surpassing Simple Aggregation: Advanced Strategies for Analyzing Contextual-Level Outcomes in Multilevel Models
Dominik Becker, Wiebke Breustedt, Christina Isabel Zuber

265 Simultaneous Feedback Models with Macro-Comparative Cross-Sectional Data
Nate Breznau

309 Blaming the Young Misses the Point: Re-assessing Young People’s Political Participation over Time Using the ‘Identity-equivalence Procedure’
Christian Schnaudt, Michael Weinhardt

335 Information for Authors

methods, data, analyses | Vol. 12(2), 2018, pp. 181-184

Editorial: Comparative Survey Analysis – Models, Techniques, and Applications

Bart Meuleman[1]*, Eldad Davidov*[2] *& Daniel Seddig*[2]

[1] KU Leuven

[2] University of Cologne and University of Zurich

The use of comparative data is of paramount importance for the understanding of societies and their change patterns. Fortunately, today more than ever, social researchers are equipped with a large number of national, international, and longitudinal comparative survey data, some of which contain repeated cross-sectional data whereas others include panel data. These data allow social scientists to test theories, generalize them across cultures, and address diverse topics of major social relevance such as attitudes toward the state and its functioning, democracy attitudes, trust in people and institutions, immigration and integration, values, and behavioral patterns, just to name a few.

However, comparative data is often characterized by a high level of complexity. The presence of multiple countries or time points can lead to complicated data structures that offer great opportunities for research but also require special methods of analysis (Van de Vijver & Leung 1997; Davidov, Schmidt, Billiet, & Meuleman 2018). This special issue is devoted to studies that demonstrate advanced techniques for analyzing comparative survey data and present applications of comparative analysis on a diverse range of topics. The special issue includes five studies. Some analyze comparative cross-sectional data, while others examine longitudinal data or a combination of both types of data. Below we provide a short overview of the studies.

The first paper, '*Modeling multiple-country repeated cross-sections: A societal growth curve model for studying the effect of the economic crisis on perceived ethnic threat*', by Bart *Meuleman, Eldad Davidov,* and *Jaak Billiet* demonstrates how to exploit the richness of comparative data which cover both multiple countries and multiple time points. It presents a novel application for cross-national time series survey data using societal growth curve modeling. While growth curve modeling has been often applied to individual data, this study shows how it may also be employed for contextual country-level data. The method is illustrated using six rounds of data from the European Social Survey (2002-2012). It inquires whether indicators of economic downturn are systematically related to increased levels of economic and cultural threat due to immigration. The societal growth curve modeling approach makes it possible to differentiate longitudinal effects from cross-sectional differences thus overcoming the weaknesses of analyses relying on single-shot cross-sectional data.

The second study, '*Demonstrating how to best examine group-based segregation: A statistical and conceptual multilevel approach*' by *Christoph Spörlein* and *Elmar Schlueter* addresses the topic of segregation between ethnic or sociodemographic groups from a comparative perspective. The authors claim that segregation has been often studied by researchers from a descriptive perspective and, consequently, these studies lack an inferential statistics approach. In their paper, they present the multilevel binomial response approach that provides a particularly flexible framework for describing and explaining segregation to better understand the role of individual- and contextual-level drivers of segregation. The authors employ three case studies using survey data from urban, national, and cross-national settings: the German urban monitoring survey, individual data from the European Labor Force Surveys in 15 EU member states, and a single wave from the German Socio-Economic Panel Study. They focus on different manifestations of ethnic and gender segregation.

The third study, '*Surpassing simple aggregation: Advanced strategies for analyzing contextual-level outcomes in multilevel models*' by *Dominik Becker*, *Wiebke Breustedt,* and *Christina Isabel Zuber* introduces two advanced analytical strategies for analyzing contextual-level outcomes in multilevel models: the multilevel SEM and the two-step approach. The authors first discuss the methodological and statistical advantages of the two approaches and then illustrate their advantages in a substantive study. Their substantive study examines the effect of citizens' support for democratic values on the persistence of democracy, drawing on data from the World Values Survey and the Quality of Government project.

The fourth paper, '*Simultaneous feedback models with macro-comparative cross-sectional data*' by *Nate Breznau* addresses advantages and limitations of comparative cross-sectional data from a different angle. The author argues that while many authors do not have access to longitudinal data, they are nevertheless interested in assessing relationships of reciprocal causality that are postulated by their theories. The paper discusses the conditions that make it possible to assess simultaneous feedback models of reciprocal causality using cross-sectional survey research. The author shows how to construct simultaneous feedback models using a structural equation modeling perspective. The method is exemplified using three commonly used software packages (MPlus, Stata, and R) and data from the International Social Survey Program covering 70 country-time points (between 1985 and 2006) to model simultaneous feedback relations between public opinion and social spending.

Finally, the fifth paper, '*Blaming the young misses the point: Re-assessing young people's political participation over time using the "identity-equivalence procedure"*', by *Christian Schnaudt* and *Michael Weinhardt* addresses the topic of construct equivalence when comparing data over time and across age groups. They suggest that construct equivalence is more important for a meaningful comparison than identical instruments that are in fact not equivalent. They exemplify the application of construct equivalence on the topic of political participation of young and older people. Specifically, they apply a procedure that they name "identity equivalence" on the measurement of political participation across three different age groups and over the time period between 2002-2014 using data from the European Social Survey.

References

Davidov, E., Schmidt, P., Billiet, J. & Meuleman, B. (2018). *Cross-cultural analysis: Methods and applications.* New York: Routledge.

Van de Vijver FJR & Leung K. (1997). *Methods and Data Analysis for Cross-Cultural Research*. London: Sage.

Acknowledgment

The second author would like to thank the University of Zurich Research Priority Program "Social Networks" for their support during work on this study and on the special issue. All authors of the Editorial would like to thank Lisa Trierweiler for the English proof.

methods, data, analyses | Vol. 12(2), 2018, pp. 185-210 DOI: 10.12758/mda.2017.10

Modeling Multiple-country Repeated Cross-sections. A Societal Growth Curve Model for Studying the Effect of the Economic Crisis on Perceived Ethnic Threat

Bart Meuleman [1], *Eldad Davidov* [2] & *Jaak Billiet* [1]

[1] University of Leuven

[2] University of Cologne and University of Zurich

Abstract

While multiple-country repeated cross-sectional datasets are increasingly available, few cross-national studies fully exploit the richness of such data. This paper contributes to the practical knowledge on statistical analysis of cross-national time series data. For that purpose, we present a novel application of a societal growth curve model (Fairbrother, 2014) analyzing the pressing question whether the economic crisis of the past years has stirred up immigration-related threat perceptions among European citizens. Concretely, we analyze six rounds of European Social Survey data (2002-2012) to investigate whether indicators of economic downturn are systematically related to increased levels of economic and cultural threat. The societal growth curve modeling approach makes it possible to set longitudinal effects apart from cross-sectional differences and thus overcomes the weaknesses of analyses relying on single-shot cross-sectional data. Our results provide evidence that growing unemployment as well as decreasing rates of economic growth instigate feelings of economic threat. Rather than affecting citizens' opinion uniformly, the economic crisis is found to have the strongest impact on economic threat among low educated people. While this study provides evidence that economic shocks affect concerns that immigration is bad for the economy, feelings of cultural threat are not affected by economic crises.

Keywords: group conflict theory, economic vs. cultural threat, societal growth curves, European Social Survey

Over the course of the last decades, cross-national data collections – such as the *European Social Survey* (ESS), the *European Values Study* (EVS), or the *International Social Survey Programme* (ISSP) – have accumulated trend data rendering it possible to monitor change in citizens' values, attitudes and behavior. These data can be characterized as cross-national repeated cross-sections: Multiple countries are observed across a time range, but at each point of observation a different cross-section of the national population is surveyed. The potential contribution of this design to social scientific insights is very large. The longitudinal aspect can help to partially overcome the well-known but crucial causality problem that single-shot cross-national studies suffer from. Cross-sectional studies can demonstrate that differences in a context variable tend to coincide with particular patterns in public opinion at a given time point. Such correlational patterns only provide a very shaky empirical foundation to make claims about causality. Cross-national trend data can provide additional insights in the temporal order of the relationship, which is a necessary (yet insufficient) condition for causality. However, according to the seminal work of Campbell and Stanley (1966; see also Shadish, Cook & Campbell, 2001), a multi-location time series design can provide interesting insights, especially when experimental manipulation is not feasible.

While multiple-country repeated cross-sections are increasingly available, knowledge on statistical tools to optimally analyze such data is limited. As a result, many current cross-national studies do not fully exploit the richness of the available data. This paper demonstrates the practical implementation of a statistical model to analyze multi-country repeated cross-sectional datasets. The second purpose of this paper is to utilize the model to analyze the effect of the economic crisis on threat due to immigration among Europeans. We do this by providing a novel application of the societal growth curve model introduced by Fairbrother (2014). This model uses multilevel techniques to estimate how a particular aggregated individual characteristic – such as ethnic threat - develops over time on the country level, and to assess whether contextual variables can explain the observed over-time developments. We apply this model to test whether the 2008 economic crisis has affected perceptions of ethnic threat among European citizens. Numerous single-shot cross-national studies have presented empirical evidence that economic

Acknowledgments

The work of the first author was supported by a Brain-be project funded by BELSPO (grant agreement BR/175/A5/PUMOMIG). The second author would like to thank the University of Zurich Research Priority Program "Social Networks" for the support during work on this study.

Direct correspondence to

Bart Meuleman, Institute for Social and Political Opinion Research, University of Leuven, Belgium

E-mail: bart.meuleman@kuleuven.be

conditions are related to prejudice, perceived threat, and anti-immigrant sentiments (for reviews, see Ceobanu & Escandell, 2010; Hainmueller & Hopkins, 2014). Studying the development of exclusionary attitudes over time in multiple countries, however, provides a more stringent test of the causal impact of economic conditions (for examples, see Meuleman, Davidov & Billiet, 2009; Semyonov, Raijman & Gorodzeisky, 2006). The societal growth curve approach allows disentangling longitudinal and cross-sectional effects of economic context.

Concretely, we employ the societal growth curve model in the current study to address the following research questions: (1) In what way has the prevalence of perceived immigrant threat in European societies evolved in the period before and after the outbreak of the economic crisis in 2008? (2) Are the observed developments in perceived threat driven by *changes* in economic conditions due to the crisis? (3) Does the crisis affect threat perceptions across the whole population, or are crisis effects instead contingent on social positions in the form of education level? To answer these questions, we analyze data from the European Social Survey (ESS) across the years 2002-2012, providing information about immigration-related threat perceptions in 28 countries before, during, and after the outbreak of the economic crisis.

The paper starts by providing the theoretical background and formulating our research hypotheses. Second, we explain how societal growth curve models can be used to test our hypotheses using multiple-country repeated cross-sections. Subsequently, we present the data and measures we use. The paper concludes with a discussion on the results of the analysis and the usefulness of the societal growth curve model.

Theoretical Background: A Dynamic Formulation of Group Conflict Theory

Group Conflict Theory (GCT) offers a framework to understand possible effects of the economic crisis on prejudice, threat perceptions and anti-immigration sentiments. The central proposition of GCT is that negative attitudes toward outgroups – such as immigrants and ethnic minorities – develop as a defensive reaction of the majority group to the perception that prerogatives of the own group are threatened (Blumer, 1958; Gorodzeisky & Semyonov, 2016; Olzak, 1992; Quillian, 1995). Not only economic goods (such as well-paid jobs, affordable housing, or the scarce resources of the welfare state), but also cultural goods (such as cultural traditions or society-specific norms and values) can become the subject of intergroup competition (Stephan et al., 1998). The distinction between the different sources of threat perceptions is of crucial importance, as economic and cultural threat perceptions can differ in their antecedents (such as social class basis) as well as in their consequences (e.g., prejudice or voting behavior) (Harell et al., 2012; Lucassen & Lubbers, 2012; Sniderman, Hagendoorn & Prior, 2004).

According to GCT, majority-members' threat perceptions are influenced by contextual factors, such as economic conditions or immigrant group size (Blalock, 1967). In times of poor economic conditions, the material goods that are the object of intergroup competition become scarcer, thereby leading to an intensification of (mainly economic) threat perceptions. Furthermore, a more sizeable immigrant group implies that the native population is confronted with a larger number of competitors, again causing intergroup competition to become stronger. Several empirical studies have confirmed that anti-immigration attitudes are more widespread in adverse economic contexts (Quillian, 1995; Schneider, 2008; Semyonov et al., 2006) with high levels of ethnic diversity (Lahav, 2004; Quillian, 1995; Scheepers, Gijsberts & Coenders, 2002; Schneider, 2008), although these effects could not always be replicated (Sides & Citrin, 2007). A serious limitation that can be often observed in this body of research is its reliance on cross-sectional data sources (Hainmueller & Hopkins, 2014). However, the finding that international differences in economic performance coincide with variations in public opinion at a given time point hardly proves that economic downturns may be a cause of threat perceptions. After all, numerous other variables – such as the immigration history of a country, the broader political climate, the media, or the implemented migration and integration policies – might intervene in the relationship between economy and public opinion (Schlueter, Meuleman & Davidov, 2013).

A dynamic reformulation of GCT (Coenders & Scheepers, 1998; Meuleman et al., 2009) instead proposes to study how attitude *changes* are driven by *changes* in the actual level of competition. The theoretical rationale for this focus on changes is that sudden shifts in economic prosperity or immigrant presence could have more substantial effects on public opinion than high but stable levels of actual competition (Hopkins, 2010). Sudden changes affect labor, housing, and other markets more strongly than slow-paced evolutions (Olzak, 1992) and usually receive wide media coverage (Schlueter & Davidov, 2013; McLaren, Boomgaarden & Vliegenthart, 2017). A crucial methodological advantage of focusing on longitudinal changes is that it offers a more stringent test of the causal relationships articulated in the GCT.

The – relatively few – empirical studies using a dynamic approach often support the propositions derived from GCT. Economic downturns were found to instigate threat perceptions and anti-immigrant attitudes in the United States (Quillian, 1996), Canada (Wilkes & Corrigall-Brown, 2011; Wilkes, Guppy & Farris, 2008), Germany (Coenders & Scheepers, 2008), and the Netherlands (Coenders & Scheepers, 1998; Coenders et al., 2008). Also, studies combining a cross-national and longitudinal perspective confirm the role of economic conditions (Semyonov et al., 2006; Meuleman et al., 2009; Kuntz et al., 2017). Pichler (2010) furthermore demonstrates that economic conditions can also alter the mechanisms through which threat perceptions are formed. During periods of unfavorable economic conditions economic concerns come to the fore in the formation of threat perceptions, while cultural concerns are suppressed.

A limitation of existing studies is that they span periods with only relatively small economic fluctuations. Yet, the recent economic turmoil might be conceived as a new critical juncture that sets in motion different mechanisms, compared to those active during more modest economic fluctuations (Billiet, Meuleman & De Witte, 2014; Semyonov et al., 2006). Little is known about the impact of a serious economic crisis. This study therefore tests whether the economic downturn Europe has been experiencing in the aftermath of the 2008 financial crisis has affected economic and cultural threat perceptions among majority-group citizens. Based on GCT, we expect that *threat perceptions have increased in Europe since the beginning of the crisis in 2008* (Hypothesis 1) and that *changes in threat perceptions in European countries are related to country-level changes in economic conditions* (Hypothesis 2). Furthermore, building on Pichler's (2010) argument on the shifting foundations of threat perceptions, we expect that *indicators of the economic context will have a stronger impact on economic than on cultural threat perceptions* (Hypothesis 3). Finally, the individual-level component of GCT suggests that the threat-inducing effect of the crisis might be stronger among individuals in social-structurally vulnerable positions in the form of low education levels, whereas there is no such effect among those who are highly educated (as a proxy for being well off). This would, in other words, imply that *the effect of the crisis on threat perceptions interacts (negatively) with education* (Hypothesis 4).

Modeling Multiple-Country Repeated Cross-sections: Societal Growth Curves

The aforementioned hypotheses can be tested by means of multiple-country repeated cross-sectional data, that is, data consisting of several countries that are observed at different time points, by surveying a large number of individuals. Such data contain a three-level hierarchical structure, with countries at the highest level, country-years at the middle level, and individuals at the lowest level. This nested structure can be taken into account by fitting a societal growth curve model (Fairbrother, 2014) that estimates how an individual characteristic evolves over time within countries - see equation (1).

$$Y_{itj} = \beta_0 + \beta_1 Time_{tj} + v_{1j} Time_{tj} + v_{0j} + u_{0tj} + e_{itj} \quad (1)$$
$$with\, e_{itj} \sim N\left(0, \sigma_e^2\right)$$

$$u_{0tj} \sim N\left(0, \sigma_u^2\right)$$

$$\nu_{0j} \sim N\left(0, \sigma_{\nu 0}^2\right)$$

$$\nu_{1j} \sim N\left(0, \sigma_{\nu 1}^2\right)$$

Y_{itj} represents a measured characteristic (e.g. perceived threat) for an individual i, surveyed at time point t in country j. β_0 is the grand intercept in this model, referring to the predicted level of Y at the beginning of the time series averaged across all countries. By including the variable 'time' as a fixed effect at the second level (country-years), the overall evolution of the dependent variable Y is modeled, which is an essential feature of the growth curve approach. In equation (1), the time effect is linear (with an effect parameter β_1), but the model can be extended in a straightforward way to include more complex functional forms of growth. Random effects for the intercept (ν_{0j}) and the slope (ν_{1j}) of the growth curve are included to accommodate the country specificity of threat developments over time, that is, how the growth curve in a specific country deviates from the average developmental pattern. The model also contains random components at the middle (u_{otj}) and lowest (e_{itj}) levels. u_{otj} reflects how country-years deviate from the country-specific growth curve. e_{itj} captures the individual-level residuals. This approach shows similarity to conventional multilevel growth curve models for panel data (e.g. Andreß, Golsch & Schmidt, 2013). The main difference is that the occurrence of repeated measurements is not at the level of individuals, but rather at the level of the countries. As a consequence, the intercepts of the growth curve are situated at the level of the country-years (level 2), and the intercept variation is captured by its variance component ν_{0j}. The slope of the growth curve is estimated by the linear effect of the time variable, the slope variation is absorbed in its variance component ν_{1j}. As such, the societal growth curve model is essentially a classical two-level growth curve model for countries, with an additional layer of individuals underneath.

One could add to this baseline model individual-level as well as contextual predictors. Of crucial importance is that the societal growth curve approach makes it possible to partition the impact of contextual variables into a cross-sectional and a longitudinal component. This decomposition takes place by simultaneously including a time-invariant (i.e., the average over the complete time series) and a time-varying component (the year-specific deviation of that average) of the contextual variables into the models (Fairbrother, 2014; this decomposition is similar to disentangling between- and within-cluster covariate effects in clustered data – see Neuhaus & Kalbfleisch, 1998).

Take a contextual variable Z_{tj} that varies across countries as well as time points (e.g., the unemployment rate). Time-invariant component $Z_{\bullet j}$ equals the value of this contextual variable for a particular country averaged over the whole observed time series (e.g., the average unemployment rate of a specific country between 2001 and 2012). The parameter for this time-invariant component captures the cross-sectional relationship between context and threat levels, irrespective of changes over time. The time-varying component is calculated as the deviation of the observed value at a specific time point from the country average over the whole time series (Z_{tj} - $Z_{\bullet j}$). The parameter for the time-varying component describes longitudinal relationships, that is, how variations in perceived threat over time within countries (from their longitudinal average) are associated with changes in a contextual variable. Because $Z_{\bullet j}$ and (Z_{tj} - $Z_{\bullet j}$) are included simultaneously in the model, the parameter for the time-varying component reflects the pure longitudinal effect, controlling for its average over the whole time series. If there is a causal impact of a particular context variable, its longitudinal effect should be different from zero.

Finally, cross-level interactions between the longitudinal variations of contextual variables and individual characteristics can be included to investigate whether the growth curve components (intercept and slope) vary across different categories of individuals.

Materials and Methods

Dataset: European Social Survey, 2002-2012

We analyze data from a time series of six rounds of the European Social Survey (ESS), spanning the period before and after the crisis (2002-2012). This multi-location time-series design is one of the strongest alternatives when experimental manipulation is not feasible, under the condition that the event that should bring about change in the time series (the quasi treatment) is well specified a priori (which is the case here) (Campbell & Stanley, 1966: 38; see also Shadish, Cook & Campbell, 2001). The logic behind it is that it is unlikely that particular quasi-experimental treatments are followed by an outcome change in multiple locations, if the effect is not causal.

Since the focus is on change, we include only countries that participated in at least two ESS rounds. Our dataset comprises 28 countries with a total of 137 country-year combinations. In all countries, strict probability samples of the resident population aged 15 years and older were drawn. Because we are interested in the attitude patterns among members of the majority population, respondents who were born outside the country, who have a foreign nationality, or who consider themselves as a member of an ethnic minority group are removed from the sample (see also Sarrasin, Green, Fasel & Davidov, 2015). The total sample size equals 228,331 individuals (for sample sizes per country and year and country abbreviations, see Appendix 1).

Measurements

Dependent variables – The ESS core module contains two items that were designed to measure economic and cultural threat perceptions.[1] Respondents are invited to position themselves on an 11-point scale of which the endpoints refer to perceiving immigration as a disadvantage or as an advantage for the economy ('Would you say it is generally bad or good for [country]'s economy that people come to live here from other countries?') or the cultural life ('Would you say that [country]'s cultural life is generally undermined or enriched by people coming to live here from other countries?'). The scales are reversed, so that 0 indicates low and 10 high threat. While these items have been used as indicators of a single concept of general group threat in previous research (Sides & Citrin, 2007), we analyze them separately to render the difference between economic and cultural sources of threat visible (for a similar approach, see Pichler, 2010).[2] This approach is justified by the fact that both items contain – especially at the individual and country-year level – considerable unique information. The correlation between economic and cultural threat equals 0.60 at the individual level, 0.71 at the country-year level, and 0.83 at the country level, implying that the two items share 36.0, 50.1, and 69.3 percent of their variance at these respective levels. These unique components allow sufficient room for differential effects of individual as well as contextual predictors (see below).

Contextual predictor variables – All contextual variables were retrieved from the Eurostat website (http://ec.europa.eu/eurostat). The economic context is captured by means of the real GDP growth rate (Eurostat indicator *nama_gdp_k*) and the harmonized unemployment rate (Eurostat indicator *une_rt_a*). Changes in immigrant group size are measured by the inflow of foreign immigrants (Eurostat indicator *migr_imm1ctz*) per capita. We include the time-invariant as well as the time-varying components of these contextual variables. Concretely, we average contextual information over two years to indicate the time-varying component referring to a specific time point (e.g., the average unemployment rate of 2001 and 2002 is taken to predict threat perceptions in the 2002 survey). This choice reflects

1 The core module of the ESS contains a third item measuring immigration-related group threats (ESS item *imwbcnt*). Because the wording of this item is very general and does not refer to specific sources of threat, we do not include it in the analysis.

2 The use of single items instead of multi-item batteries makes it difficult to assess the reliability, validity, and cross-cultural comparability of the measurements. To get an indication of the measurement quality, we performed multiple group confirmatory factor analysis (Davidov et al., 2014) on the three threat items included in the ESS across our 137 country-year combinations. Partial measurement equivalence could be established for all countries but Ireland (the output can be obtained from the first author upon request). As a result, the data allow us to conduct meaningful comparisons across all countries and time points. To rule out that the outlier Ireland biases our conclusions, we re-estimated all our models excluding the Irish data as a robustness check, but the effects of the economic context remained unchanged.

that the impact of economic contexts may be lagged. The time-invariant component is the average across the whole time series (2002-2012).

Individual-level predictor variables – In order to control for compositional differences – that is, the fact that European populations have a different composition in terms of several individual characteristics - we include a series of variables capturing social-structural positions and cultural dispositions that were shown to be relevant in previous research (e.g. Coenders & Scheepers, 1998; Meuleman, Davidov & Billiet 2009; Meuleman, Abts, Slootmaeckers, & Meeusen, 2018; Semyonov, Raijman & Gorodzeisky 2006). The social-structural variables are *gender*, *age*, number of years of *education* completed, *degree of urbanization* (from 1 = countryside to 5 = big city), *employment status* (distinguishing self-employed, higher service class, white collar, blue collar, unemployed, retired, in education, doing housework, disabled, and other) and *subjective income*. The latter variable is used as a proxy for the household income and is operationalized by the individual assessment of whether one finds it difficult or comfortable to live on the present income (1 = very difficult; 2 = difficult; 3 = coping; 4 = living comfortably). Based on previous literature, we expect people in socially vulnerable positions, that is with lower education and lower subjective income, the unemployed and the low-skilled workers to feel more threatened by immigrants. Furthermore, older individuals are expected to be more negative toward immigrants (e.g., Hercowitz-Amir, Raijman & Davidov 2017; Meuleman, Davidov & Billiet, 2009; Semyonov, Raijman & Gorodzeisky, 2006).

Religious involvement is the mean of items measuring subjective religiosity (ESS item *rlgdgr*), attendance of religious services (*rlgatnd*) and frequency of praying (*pray*). *Political orientation* is measured by self-placement on a left (0) to right (10) scale. This scale was categorized into three groups, namely, left (scores 0-4), center (5), and right (6-10). To handle the considerable nonresponse of this item, we added a fourth category for the missing values. Secular persons as well as left-leaning individuals are assumed to express lower levels of perceived ethnic threat (see, e.g., Hercowitz-Amir et al., 2017).

Descriptive statistics for the variables are displayed in Appendix 2.

Statistical Modeling

The random effect models are estimated by means of the MIXED procedure of SAS 9.3, using a restricted maximum likelihood estimation method. To obtain standard errors that are robust against deviations of the distributional assumptions of the random effects (such as non-normality), we furthermore used the "sandwich estimator" (Verbeke & Molenberghs, 2000: 87ff). All analyses are weighted to correct for cross-national differences in sampling design (*dweight*). All continuous individual-level predictor variables were centered around their grand mean prior to the analysis. Apart from political orientation – where a separate category for the missing values is created – we applied listwise deletion to deal with the item non-response. The amount of missing values in the data was quite limited and lower than 5% on average ranging between 4.6% for the variable economic threat and 0.1% for gender. Therefore, we do not expect that using listwise deletion distorts our conclusions (see Schafer & Graham, 2002).

Results

Trends in Perceived Threat, 2002-2012

Before presenting the societal growth curves, we explore the development in threat perceptions over the period 2002-2012. Considerable cross-country differences can be observed in the level of perceived *economic* threat (see Figure 1), ranging from as low as 3.36 (Luxemburg, 2002) to as high as 7.22 (Cyprus, 2012) (on a scale from 0 to 10). These differences follow regional patterns, with the lowest levels of economic threat in Northern Europe and the highest scores in Eastern and Southern Europe. Longitudinal developments within countries appear to be smaller than between-country differences. The most notable change is observed in Ireland, where economic threat shifts from 4.04 (2006) to 5.85 (2010). Progression of economic threat is patterned along regional lines as well. In the Nordic countries, which already displayed comparatively low threat in 2002, economic threat perceptions tend to stabilize or even diminish. In Southern Europe, by way of contrast, a clear upward trend is notable. It is revealing to observe that between 2008 (the outbreak of the financial crisis) and 2010 (when its impact on the economy was becoming clear), economic threat perceptions became more prevalent in 20 countries, while they became weaker in 3 countries only (see also Kuntz et al., 2017).

Regarding cultural threat (Figure 2), the specific position of Scandinavian countries becomes even more distinct. Northern Europeans perceive substantially less cultural threat compared to citizens in Western, Eastern, and Southern Europe. Importantly, longitudinal changes in cultural threat are less outspoken than in the case of economic threat. At least during our time window of observation, cultural threat perceptions seem to be a more stable phenomenon, while economic threat perceptions tend to fluctuate substantially.

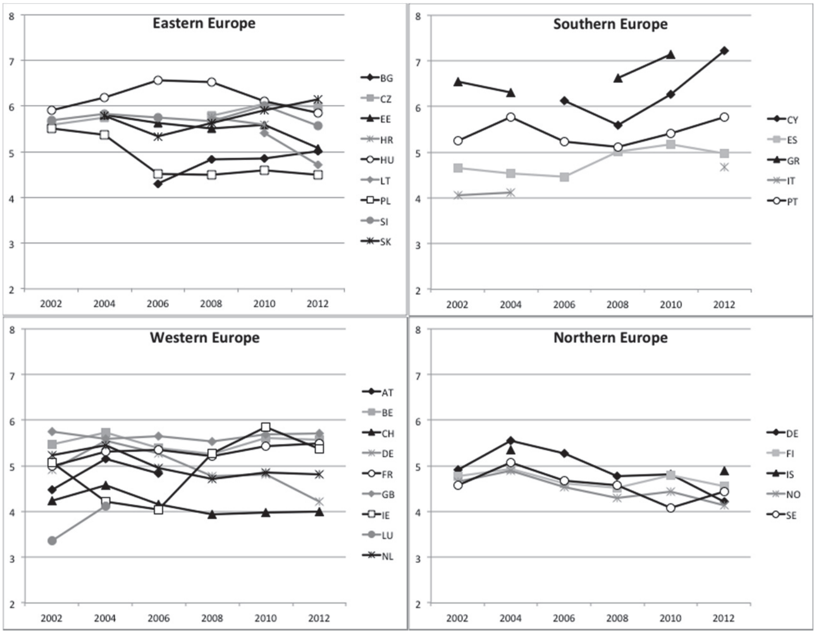


Figure 1 Development of perceived economic threat in 28 countries (by region) – 2002-2012

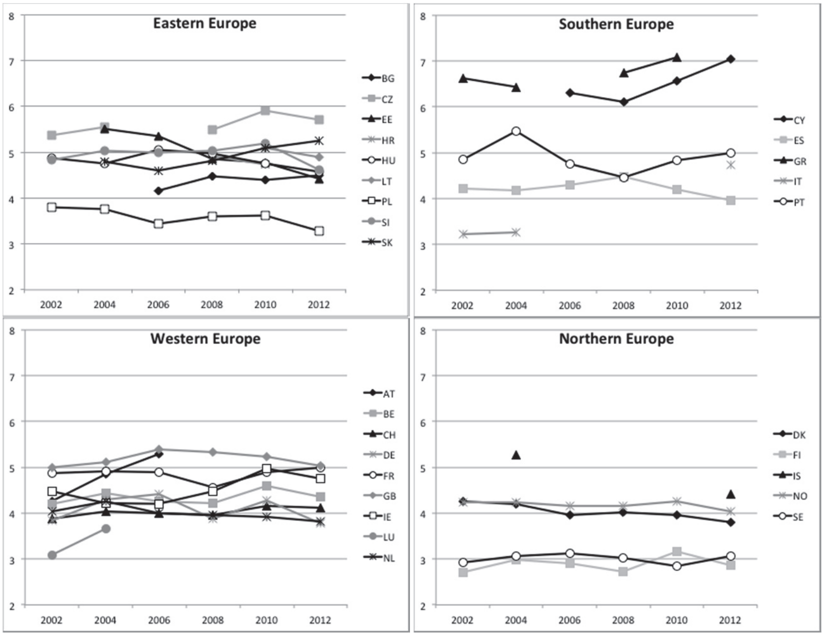


Figure 2 Development of perceived cultural threat in 28 countries (by region) – 2002-2012

Societal Growth Curves: The Longitudinal Impact of Economic Conditions

To examine the effects of economic conditions on threat perceptions, we estimate a series of societal growth curve models for economic and cultural threat (see Tables 1 and 2). An empty three-level model (not shown) indicates that economic and cultural threat perceptions vary significantly across individuals, country time points combinations as well as countries. The lion's share of the total variation can be attributed to the individual level. Variations of threat between countries (7.5% of the total variance for economic threat vs. 12.9% for cultural threat) are considerably larger than longitudinal variations of threat within countries. Notably, the longitudinal variation of economic threat (2.0%) is more than double than that of cultural threat (0.9%).

Models 1E (Table 1) and 1C (Table 2) estimate growth curves by including time as a predictor. A linear time trend combined with a dummy for 2010 (picking up an additional change in 2010 over and above the linear process) provides the most appropriate description of the data. For both forms of threat, the linear time effect is insignificant, but does have significant random slope variation. This means that, *on average* across all countries, threat perceptions remain stable between 2002 and 2012; the linear trend does vary cross-nationally, however, with increases in some countries and decreases in others. One particular ripple disturbs the linear development of threat perceptions. The dummy for 2010 has a significant and positive effect. In 2010 (i.e., following the outbreak of the financial crisis), economic and cultural threat perceptions were respectively 0.116 and 0.120 units higher than what is expected based on the general time trend. This pattern confirms that immigrant-related threat perceptions have increased across Europe after the 2008 crisis (supporting Hypothesis 1), although the magnitude of the increase should not be overrated. Furthermore, the 2010 increase in threat perceptions was instantaneous and had receded by 2012.

Indicators of the economic context as well as individual characteristics are added in Models 2E and 2C. Economic and cultural threat perceptions are – to a large extent but not completely – driven by the same set of individual predictors. As expected by theories of ethnic competition, threat perceptions are most outspoken among individuals with a lower socioeconomic status. Fewer years of education and a lower (subjective) income seem conducive towards increased threat perceptions. Concerning employment status, the highest levels of threat perceptions are observed among blue collar workers, followed by persons who are unemployed, retired, disabled, or homemakers. Members of the higher service class and those in education feel least threatened. Furthermore, also persons living in a rural environment express higher levels of economic and cultural threat. Consistent with previsous research (e.g. Semyonov, Raijman & Gorodzeisky, 2006), political orientation is among the strongest predictors of perceived threat: left-leaning individuals feel culturally as well as economically less threatened. Apart from these similarities, three individual variables have a differential impact. Males feel economically less threatened than females, while no gender gap is present for cultural threat. Furthermore, the highest levels of cultural threat are found among respondents between the ages of 55 and 74 years, while this age group does not deviate from the reference category (aged 45-54 years) on economic threat. Finally, religiosity has a small tempering effect on economic threat but shows no significant relationship with cultural threat.

To find out whether *changes* in the economic context affect threat perceptions, Models 2E and 2C include the country time-invariant (cross-sectional) and the time-varying (longitudinal) components of two economic variables, namely, the unemployment rate and the real GDP growth. The longitudinal components of unemployment and economic growth have a significant impact on feelings of *economic* threat. In times of rising unemployment rates and plummeting growth rates, citizens' anxieties that immigration poses a threat to the national economy gain momentum (supporting Hypothesis 2). These longitudinal effects of economic context are substantial. Spain, for example, experienced an increase in unemployment rate of 12.4 percentage points and a drop in economic growth of 3.8 percentage points between 2005-6 (the 3rd ESS round) and 2010-11 (the 5th ESS round). Model 2E predicts that the combination of these economic shocks increased economic threat perceptions across the whole Spanish population with more than 0.6 points, which implies a considerable increase. It is of crucial importance to reiterate that these parameters refer to longitudinal effects, capturing the impact that national economic conditions at particular time points have on the evolution of threat perceptions within countries. At the same time, no significant cross-sectional relationships between the average country levels of economic context and economic threat are detected. Model 2E explains 7.8% of the individual variation, 42.7% of the variation between country-years and 25.1% of the between-country differences in economic threat. The model is thus quite successful in explaining why a country's level of economic threat is higher at particular time points than in other years. Note that the effect of the dummy for 2010 has become insignificant, indicating that the high levels of economic threat in that particular year are indeed driven by economic changes.

Whereas economic conditions shape the development of perceived economic threat, no such contextual effects are found for cultural threat. The idea that immigration threatens the nation's cultural life is not only relatively stable over time, but also completely detached from economic changes. Crisis-induced threat perceptions seem to be limited to concerns about economy, and do not generalize to the cultural realm. This finding is in line with Hypothesis 3. For cultural threat, Model

2C explains 22.7%, 37.3%, and 8.0% of the variance of the dependent variable at the country, country-time, and individual levels respectively.

One of the specific features of the societal growth curve approach is that contextual variables are decomposed into a cross-sectional and a longitudinal component. In order to scrutinize the similarities and differences with the classical approach –that is, including the raw context variables, without decomposition- we additionally estimated models in which only the unemployment rates and GDP growth scores in the year of the survey are included.[3] We find that the effects of unemployment rate (on economic threat: 0.0351; on cultural threat: 0.0043) and GDP growth (on economic threat: -0.0530; on cultural threat: -0.0031) are very similar to the longitudinal effects found in Models 2E and 2C. This similarity is however particular for the current analysis. It is most likely a result of the fact that the cross-sectional effects per se in our growth models are quite small and insignificant. This may not always be the case, however. In some cases the cross-sectional component of a country score may have an effect on the dependent variable that is stronger or even opposite compared to the effect of the longitudinally varying component. Without decomposition, the estimated context effect is a mixture between the cross-sectional and the longitudinal effect. If both effects are considerable and different, omitting the decomposition can lead to incorrect conclusions.

As an additional robustness check, we re-estimated the effect of economic conditions on economic and cultural threat respectively, controlling for the inflow of foreign immigrants (per capita).[4] Neither the longitudinal nor the cross-sectional components of foreign immigration are related to either economic or cultural threat perceptions. The most important conclusion from this additional model is that the longitudinal effects of the economic variables unemployment and economic growth on economic threat remain significant, and are thus not driven by a possible connection between migration movements and the severity of the economic crisis.

3 We would like to thank an anonymous reviewer for this suggestion. The full results are not shown here, but can be obtained from the first author.

4 The full results are not shown here, but can be obtained from the first author. This control variable was only included in this stage of the modeling process because the migration flow statistics contain several missing values and lead to the exclusion of the following country-years: FR 2002; FR 2004; GR 2002; GR 2004; IS 2004; PT 2002; PT 2004; PT 2006).

Table 1 Societal growth curve models for economic threat

	Model 1E		Model 2E		Model 3E	
Fixed effects - indiv. level	Par. Est.	SE	Par. Est.	SE	Par. Est.	SE
Intercept	5.155	(0.124)***	5.195	(0.287)***	5.650	(0.348)***
Time	-0.006	(0.025)	-0.031	(0.016)	-0.040	(0.025)
Dummy: time-point 2008	0.116	(0.055)*	-0.122	(0.068)	-0.159	(0.087)
Gender						
male			-0.264	(0.028)***	-0.261	(0.027)***
female (ref.cat.)						
Age category						
16-24 years			0.155	(0.056)**	0.141	(0.055)*
25-34 years			0.103	(0.031)***	0.100	(0.030)***
35-44 years			0.049	(0.021)*	0.050	(0.020)*
45-54 years (ref.cat.)						
55-64 years			0.023	(0.033)	0.025	(0.032)
65-74 years			0.023	(0.058)	0.029	(0.057)
75 years and over			0.047	(0.064)	0.044	(0.064)
Education			-0.099	(0.006)***	-0.101	(0.006)***
Activity status						
Self-employed			-0.300	(0.044)***	-0.295	(0.044)***
Higher service class			-0.632	(0.056)***	-0.614	(0.054)***
White collar			-0.393	(0.039)***	-0.389	(0.038)***
Blue collar (ref.cat.)						
Unemployed			-0.102	(0.040)*	-0.103	(0.040)*
Retired			-0.183	(0.033)***	-0.187	(0.033)***
In education			-0.764	(0.045)***	-0.758	(0.044)***
Doing housework			-0.235	(0.037)***	-0.219	(0.034)***
Disabled			-0.022	(0.049)	-0.032	(0.050)
Other			-0.405	(0.066)***	-0.391	(0.065)***
Subjective income			-0.300	(0.013)***	-0.307	(0.012)***
Urbanization			-0.068	(0.008)***	-0.069	(0.007)***
Religious involvement			-0.030	(0.009)***	-0.028	(0.009)**
Left-right placement						
Left (ref.cat.)						
Centre			0.370	(0.053)***	0.357	(0.053)***
Right			0.308	(0.082)***	0.300	(0.082)***
Missing			0.600	(0.060)***	0.593	(0.060)***

Table 1 continued

	Model 1E		Model 2E		Model 3E	
Fixed effects - indiv. level	Par. Est.	SE	Par. Est.	SE	Par. Est.	SE
Fixed effects - context variables						
Unemp. rate - Longitudinal			0.035	(0.012)**	0.033	(0.016)*
Unemp. - Cross-sectional			0.049	(0.042)	0.010	(0.051)
GDP growth - Longitudinal			-0.052	(0.015)***	-0.050	(0.023)*
GDP growth - Cross-sectional			-0.115	(0.106)	-0.180	(0.134)
Education x Unemp. rate - Longit.					-0.004	(0.001)***
Random effects						
Level 3: Var. country intercept	0.366	(0.114)***	0.326	(0.102)***	0.311	(0.155)**
Level 3: Var. slope time	0.010	(0.005)*	0.003	(0.003)	0.005	(0.004)
Level 2: Var. country-year intercept	0.083	(0.013)***	0.065	(0.011)***	0.065	(0.023)***
Level 2: Var. slope education					0.000	(0.000)***
Level 1: Residual variance	5.224	(0.016)***	4.817	(0.015)***	4.787	(0.015)***
Deviance	941487.7		924925.3		924479.9	

*p<.05; **p<.01; ***p<.001; $N_{\text{individuals}}$=205,759, $N_{\text{country-years}}$=137, $N_{\text{countries}}$=28

Table 2 Societal growth curve models for cultural threat

	Model 1C		Model 2C		Model 3C	
Fixed effects - indiv. level	Par. Est.	SE	Par. Est.	SE	Par. Est.	SE
Intercept	4.488	(0.160)***	4.173	(0.336)***	5.087	(0.388)***
Time	0.006	(0.021)	0.014	(0.020)	0.002	(0.028)
Dummy: time-point 2008	0.120	(0.035)**	0.085	(0.065)	0.050	(0.098)
Gender						
male			0.057	(0.036)	0.062	(0.036)
female (ref.cat.)						
Age category						
16-24 years			0.133	(0.061)*	0.113	(0.061)
25-34 years			0.002	(0.035)	-0.002	(0.033)
35-44 years			-0.026	(0.016)	-0.025	(0.015)
45-54 years (ref.cat.)						
55-64 years			0.066	(0.028)*	0.069	(0.028)*
65-74 years			0.146	(0.054)**	0.159	(0.052)**
75 years and over			0.234	(0.059)***	0.238	(0.058)***

Table 2 continued

	Model 1C		Model 2C		Model 3C	
Fixed effects - indiv. level	Par. Est.	SE	Par. Est.	SE	Par. Est.	SE
Education			-0.103	(0.008)***	-0.105	(0.008)***
Activity status						
Self-employed			-0.228	(0.042)***	-0.219	(0.043)***
Higher service class			-0.505	(0.060)***	-0.473	(0.056)***
White collar			-0.407	(0.041)***	-0.400	(0.040)***
Blue collar (ref.cat.)						
Unemployed			-0.144	(0.051)**	-0.146	(0.050)**
Retired			-0.099	(0.036)**	-0.103	(0.036)**
In education			-0.722	(0.052)***	-0.712	(0.051)***
Doing housework			-0.199	(0.039)***	-0.176	(0.038)***
Disabled			0.000	(0.055)	-0.018	(0.053)
Other			-0.377	(0.077)***	-0.363	(0.076)***
Subjective income			-0.230	(0.015)***	-0.243	(0.015)***
Urbanization			-0.059	(0.013)***	-0.064	(0.012)***
Religious involvement			-0.014	(0.009)	-0.011	(0.009)
Left-right placement						
Left (ref.cat.)						
Centre			0.492	(0.064)***	0.468	(0.063)***
Right			0.564	(0.105)***	0.548	(0.105)***
Missing			0.682	(0.079)***	0.669	(0.078)***
Fixed effects - context variables						
Unemp. rate - Longitudinal			0.004	(0.012)	-0.012	(0.021)
Unemp. - Cross-sectional			0.033	(0.049)	-0.051	(0.052)
GDP growth - Longitudinal			-0.003	(0.012)	-0.011	(0.026)
GDP growth - Cross-sectional			-0.084	(0.138)	-0.198	(0.138)
Education x Unemp. rate - Longit.					-0.003	(0.002)
Random effects						
Level 3: Var. country intercept	0.713	(0.202)**	0.619	(0.184)**	0.673	(0.210)**
Level 3: Var. slope time	0.008	(0.004)*	0.008	(0.004)*	0.009	(0.006)
Level 2: Var. country-year intercept	0.034	(0.006)***	0.034	(0.006)***	0.083	(0.019)***
Level 2: Var. slope education					0.002	(0.000)***
Level 1: Residual variance	5.344	(0.017)***	4.917	(0.015)***	4.885	(0.015)***
Deviance	946739.0		929758.3		928820.8	

*p<.05; **p<.01; ***p<.001; $N_{individuals}$= 205,905, $N_{country\text{-}years}$=137, $N_{countries}$=28

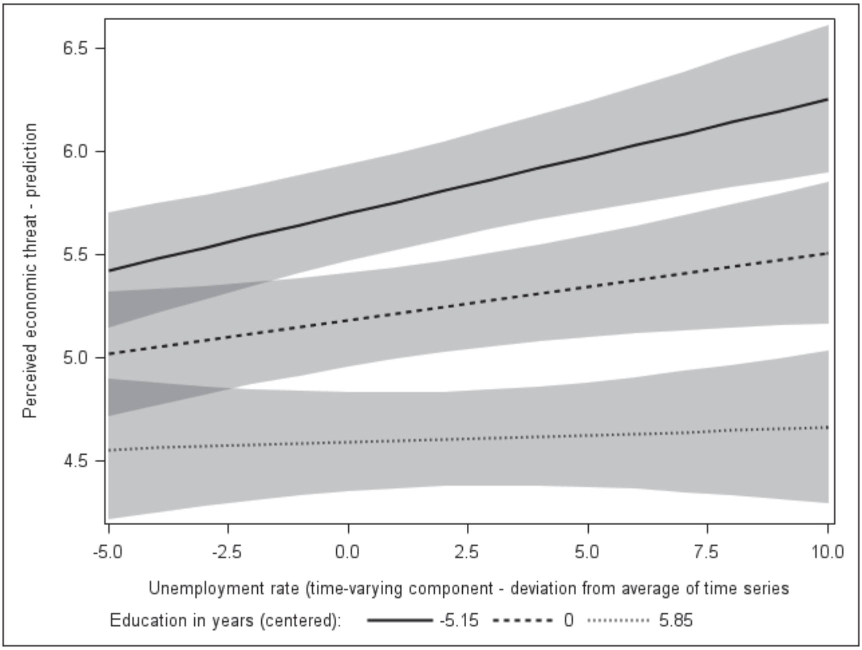
The fourth hypothesis – namely, that the longitudinal effects of the economic context are stronger among low-educated individuals – is tested in Models 3E and 3C. These models now contain a random slope for education (implying that the educational gradient of threat perceptions can vary across countries and time points) as well as a cross-level interaction effects between education and the time-varying component of the unemployment rate (testing whether the longitudinal effect of unemployment rates differs across educational groups).[5] In the case of economic threat, the longitudinal effects of unemployment are indeed different for various educational groups (see Figure 3). For an individual with an average level of education (12.5 years), represented by the middle line in Figure 3, economic threat perceptions increase by 0.033 point for every percentage point increase in unemployment rate. The negative cross-level interaction parameter (-0.004) indicates that this effect of unemployment becomes weaker as education increases. For individuals who have had 19 years of education (i.e., 6.85 years more than the average, corresponding to the 90th percentile in the dataset), the longitudinal effect of unemployment approaches zero, meaning that unemployment rates are no longer related to threat levels. For respondents with only 7 years of formal education (i.e., 5.85 years less than the average, corresponding to the 10th percentile), the impact of unemployment context is twice as strong as for the average person. This significant cross-level interaction effect shows that contextual labor market processes do not instigate economic threat perceptions uniformly across the whole population. Instead, this sociotropic source of threat seems to affect, in the first place, persons with lower education (i.e., those with a more vulnerable position in the society and the labor market), while the threat perceptions of the highly educated are more immune to the impact of labor market changes.

A similar test (not shown here) revealed that the cross-level interaction between real GDP growth rate and education is insignificant. Hereby, Hypothesis 4 is only partially supported by the data. In the case of cultural threat, none of the cross-level interactions was significant (which is not surprising given that the main effect of the economic context was insignificant for cultural threat).

In sum, this analysis reveals that economic threat perceptions have increased after the 2008 crisis (supporting Hypothesis 1), although the increase was only temporary. Furthermore, the changes in threat perceptions are driven by changes in the economic context (supporting Hypothesis 2) and are only observed for the economic component of threat (Hypothesis 3). Finally, the effects of economic conditions are more outspoken of the lower-educated individuals (Hypothesis 4).

---

5 We test the cross-level interaction for education rather than for employment status, because the latter variable is categorical which makes the estimation and interpretation of the interaction more difficult and less insightful. A similar hypothesis could in principle be tested for subjective income. However, including multiple interactions of connected variables at the same time makes the results less insightful.



This figure represents predicted levels of economic threat for various values of education (10th percentile in the highest curve, 50th percentile in the middle curve, 90th percentile in the lowest curve) and the time-varying component of the unemployment rate (full range), as well as 95% confidence bands for these predictions (the grey zone around the curves).

Figure 3 The interaction effect between education and the time-varying component of national unemployment rates

Conclusions and Discussion

The first purpose of this paper is to demonstrate the practical implementation of a statistical model to analyze multi-country repeated cross-sectional datasets. While such datasets are increasingly available, few cross-national studies optimally exploit the richness of datasets containing information on citizens surveyed in various countries and at different time points. The second purpose of this paper is to utilize the model to analyze the effect of the economic crisis on threat due to immigration among Europeans. We do this by providing a novel application of the societal growth curve model introduced by Fairbrother (2014) testing whether the 2008 economic crisis has affected perceptions of ethnic threat among European citizens. More concretely, drawing on the dynamic version of group conflict theory, the current study addressed the following three research questions: (1) In what way has the prevalence of perceived immigrant threat in European societies evolved in the period before and after the peak of the economic crisis in 2008? (2) Are the observed developments in perceived economic and cultural threat driven by crisis-related changes in economic conditions? (3) Does the crisis affect threat perceptions across the whole population, or are crisis effects instead contingent on socioeconomic status? We answered these questions by analyzing ESS data from 28 different European countries spanning the years 2002 to 2012.

Societal growth curve analysis substantiates in various ways that economic contexts shape the majority group perceptions that immigration poses a threat to the national economy. Between 2008 (just before or during the outbreak of the financial crisis) and 2010 (i.e., when the impact of the crisis on the "real economy" was becoming clear), we detected an increase – albeit short-lived – in economic threat perceptions in 20 European countries. Even more conclusive is the finding that rates of unemployment and economic growth have a longitudinal effect on economic threat perceptions: In times when unemployment rates increase and growth rates plummet, citizens' perceptions that immigrants threaten the economy become more widespread. These effects are purely longitudinal in the sense that they refer to the dynamics within countries (instead of cross-sectional differences between countries), and therefore lend strong support to the dynamic version of group threat theory. The deterioration of economic conditions in Europe indeed instigated fears that immigrants threaten economic prerogatives of the majority group, which might in turn open the door to exclusionary attitudes and discriminatory behavior. The difficult economic situation that Europe has been facing over the past years offers a breeding ground in which economic threat perceptions can easily take root. Finally, the model demonstrated that the effect of the economic crisis (i.e., increasing unemployment rates) is stronger among individuals with lower educational credentials.

The impact of economic conditions on threat perceptions is substantial, but should not be overstated and qualified in several respects. First of all, despite the fact that our analysis covered a period of unprecedented economic instability, changes in threat perceptions remain relatively limited. Differences between countries or citizens are markedly more outspoken than longitudinal variation. A severe economic shock (comparable to what a country like Spain experienced) produces an effect similar in size to the effect of social class (blue-collar workers vs. higher service class) or political orientation (left vs. right), but does not exceed the joint impact of individual-level predictors. Second, our results suggest that the economic crisis had an instantaneous effect rather than a long-lasting one. Threat perceptions did increase in the aftermath of the 2008 outbreak of the crisis, but had fallen back to pre-crisis levels by 2012. As soon as the labor market recovers and economic production takes off again, economic threat perceptions dissipate. Third, the impact of the economic crisis appears to be restricted to economic threat. Feelings of cultural threat are found to be relatively stable over time and to be completely detached from economic dynamics. At least within our window of observation, crisis-induced threat perceptions do not generalize to the idea that immigrants pose a threat to cultural life.

In sum, societal growth curve models offer promising opportunities to investigate the drivers and timing of attitude change. Further research could take this argument and method even further, for example by investigating shorter time spans, and linking public opinion to monthly instead of yearly context data. Our study shows that the societal growth curve models offer opportunities to analyze cross-national repeated cross-sections. Most importantly, by distinguishing between cross-sectional and longitudinal context effects, this approach successfully avoids the problem of weak internal validity that one faces when analyzing single-shot cross-sectional data.

References

Andreß, H. J., Golsch, K., & Schmidt-Catran, A.W. (2013). *Applied panel data analysis for economic and social surveys*. New York: Springer.

Billiet, J., Meuleman, B., & De Witte, H. (2014). The Relationship between Ethnic Threat and Economic Insecurity in Times of Economic Crisis: Analysis of European Social Survey Data. *Migration Studies*, *2*(2), 135-61.

Blalock, H.M. (1967). *Toward a Theory of Minority-Group Relations*. New York: John Wiley and Sons.

Blumer, H. (1958). Race Prejudice as a Sense of Group Relation. *Pacific Sociological Review*, *1*, 3–7.

Campbell, D.T., & Stanley, J.C. (1966). *Experimental and Experimental Quasi-Experimental Design for Research*. Boston: Houghton Mifflin Company.

Ceobanu, A.M., & Escandell, X. (2010). Comparative Analyses of Public Attitudes toward Immigrants and Immigration Using Multinational Survey Data: A Review of Theories and Research. *Annual Review of Sociology*, *36*, 309-28.

Coenders, M., & Scheepers, P. (1998). Support for Ethnic Discrimination in the Netherlands 1979–1993: Effects of Period, Cohort and Individual Characteristics. *European Sociological Review*, *14*(4), 405–22.

Coenders, M., & Scheepers, P. (2008). Changes in Resistance to the Social Integration of Foreigners in Germany 1980-2000: Individual and Contextual Determinants. *Journal of Ethnic and Migration Studies*, *34*, 1-26.

Coenders, M., Lubbers, M., Scheepers, P., & Verkuyten, M. (2008). More than Two Decades of Changing Ethnic Attitudes in the Netherlands. *Journal of Social Issues*, *64*(2), 269-85.

Davidov, E., Meuleman, B., Cieciuch, J., Schmidt, P., & Billiet, J. (2014). Measurement Equivalence in Cross-National Research. *Annual Review of Sociology*, *40*, 55-75.

Fairbrother, M. (2014). Two Multilevel Modeling Techniques for Analyzing Comparative Longitudinal Survey Datasets. *Political Science Research and Methods*, *2*(1), 119-40.

Gorodzeisky, A., & Semyonov, M. (2016). Not Only Competitive Threat But Also Racial Prejudice: Sources of Anti-Immigrant Attitudes in European Societies. *International Journal of Public Opinion Research*, *28*(3), 331-354. doi: 10.1093/ijpor/edv024

Hainmueller, J. & Hopkins, D.J. (2014). Public Attitudes Toward Immigration. *Annual Review of Political Science*, *17*, 225-49.

Harell, A., Soroka, S., Iyengar, S., & Valentino, N. (2012). The Impact of Economic and Cultural Cues on Support for Immigration in Canada and the United States. *Canadian Journal of Political Science*, *45*(3), 499-530.

Hercowitz-Amir, A., Raijman, R., & Davidov, E. (2017). Host or hostile? Attitudes towards asylum seekers in Israel and in Denmark. *International Journal of Comparative Sociology,* online first. DOI: 10.1177/0020715217722039

Hopkins, D. J. (2010). Politicized places: Explaining where and when immigrants provoke local opposition. *American political science review, 104*(1), 40-60.

Kuntz, A., Davidov, E., & Semyonov, M. (2017). The Dynamic Relations between Economic Conditions and Anti-Immigrant Sentiment: A Natural Experiment in Times of the European Economic Crisis. *International Journal of Comparative Sociology*. DOI: 10.1177/0020715217690434

Lahav, G. (2004). *Immigration and Politics in the New Europe: Reinventing Borders*. Cambridge: Cambridge University Press.

Lucassen, G., & Lubbers, M. (2012). Who Fears What? Explaining Far-Right-Wing Preference in Europe by Distinguishing Perceived Cultural and Economic Threats. *Comparative Political Studies*, *45*(5), 547-75.

McLaren, L., Boomgaarden, H., & Vliegenthart, R. (2017). News Coverage and Public Concern About Immigration in Britain. *International Journal of Public Opinion Research*. doi: 10.1093/ijpor/edw033

Meuleman, B., Abts, K., Slootmaeckers, K., & Meeusen, C. (2018). Differentiated threat and the genesis of prejudice. Group-specific antecedents of homonegativity, Islamophobia, anti-Semitism and anti-immigrant attitudes. *Social Problems*.

Meuleman, B., Davidov, E., & Billiet, J. (2009). Changing Attitudes toward Immigration in Europe, 2002-2007: A Dynamic Group Conflict Theory Approach. *Social Science Research*, *38*, 352-65.

Neuhaus, J.M., & Kalbfleisch, J.D. (1998). Between- and Within-Cluster Covariate Effects in the Analysis of Clustered Data. *Biometrics*, *54*(2), 638-45.

Olzak, S. (1992). *Dynamics of Ethnic Competition and Conflict*. Stanford: Stanford University Libraries.

Pichler, F. (2010). Foundations of Anti-immigrant Sentiment: The Variable Nature of Perceived Group Threat Across Changing European Societies, 2002-2006. *International Journal of Comparative Sociology*, *51*(6), 1-25.

Quillian, L. (1995). Prejudice as a Response to Perceived Group Threat: Population Composition and Anti-immigrant and Racial Prejudice in Europe. *American Sociological Review*, *60*, 586–611.

Quillian, L. (1996). Group Threat and Regional Changes in Attitudes toward African Americans. *American Journal of Sociology*, *102*, 816–60.

Sarrasin, O., Green, E. G. T., Fasel, N., & Davidov, E. (2015). Does survey respondents' immigration background affect the measurement and prediction of immigration attitudes? An Illustration in Two Steps. *International Journal of Public Opinion Research, 27(2),* 264-276. DOI: 10.1093/ijpor/edu015

Schafer, J. L., & Graham, J. W. (2002). Missing Data: Our View of the State of the Art. *Psychological Methods,* 7(2), 147–177.

Scheepers, P., Gijsberts, M., & Coenders, M. (2002). Ethnic Exclusionism in European Countries. Public Opposition to Grant Civil Rights to Legal Migrants as a Response to Perceived Ethnic Threat. *European Sociological Review*, *18*, 1-18.

Schlueter, E., & Davidov, E. (2013). Contextual Sources of Perceived Group Threat: Negative Immigration-Related News Reports, Immigrant Group Size and Their Interaction, Spain 1996–2007. *European Sociological Review*, *29*(2), 179-91.

Schlueter, E., Meuleman, B., & Davidov, E. (2013). Immigrant Integration Policies and Perceived Group Threat: A Multilevel Study of 27 Western and Eastern European Countries. *Social Science Research*, *42*(3), 670-82.

Schneider, S.L. (2008). Anti-Immigrant Attitudes in Europe: Outgroup Size and Perceived Ethnic Threat. *European Sociological Review*, *24*(1), 53-67.

Semyonov, M., Raijman, R., & Gorodzeisky, A. (2006). The Rise of Anti- Foreigner Sentiment in European Societies, 1988-2000. *American Sociological Review*, *71*, 426-49.

Shadish, W.R., Cook, T. D., & Campbell, D. T. (2001). *Experimental and Quasi-Experimental Designs for Generalized Causal Inference*. Boston: Houghton Mifflin

Sides, J., & Citrin, J. (2007). European Opinion about Immigration: The Role of Identities, Interests and Information. *British Journal of Political Science*, *37*(3), 477-504.

Sniderman, P., Hagendoorn, L., & Prior, M. (2004). Predisposing Factors and Situational Triggers: Exclusionary Reactions to Immigrant Minorities. *American Political Science Review*, *98*(1), 35-49.

Verbeke, G., & Molenberghs, G. (2000). *Linear mixed models for longitudinal data*. Springer New York.

Wilkes, R., & Corrigall-Brown, C. (2011). Explaining Time Trends in Public Opinion: Attitudes towards Immigration and Immigrants. *International Journal of Comparative Sociology*, *52*(1), 79-99.

Wilkes, R., Guppy, N., & Farris, L. (2008). 'No Thanks, We're Full': Individual Characteristics, National Context, and Changing Attitudes toward Immigration. *International Migration Review*, *42*(2), 302-89.

Appendices

Appendix 1 Sample sizes per country and year

	Round 1 2002	Round 2 2004	Round 3 2006	Round 4 2008	Round 5 2010	Round 6 2012	Total
Austria (AT)	1,973	2,023	2,198	- -	- -	- -	6,194
Belgium (BE)	1,700	1,574	1,611	1,535	1,473	1,565	9,458
Bulgaria (BG)	- -	- -	1,179	1,816	1,978	1,844	6,817
Switzerland (CH)	1,610	1,671	1,402	1,338	1,094	1,079	8,194
Cyprus (CY)	- -	- -	932	1,105	1,000	985	4,022
Czech Republic (CZ)	1,278	2,851	- -	1,937	2,281	1,929	10,276
Germany (DE)	2,648	2,575	2,619	2,459	2,686	2,597	15,584
Denmark (DK)	1,417	1,404	1,404	1,491	1,453	1,518	8,687
Estonia (EE)	- -	1,395	958	1,147	1,383	1,714	6,597
Spain (ES)	1,616	1,489	1,682	2,305	1,660	1,633	10,385
Finland (FI)	1,924	1,977	1,824	2,118	1,797	2,079	11,719
France (FR)	1,314	1,621	1,762	1,861	1,532	1,715	9,805
Great Britain (GB)	1,796	1,662	2,086	2,037	2,070	1,946	11,597
Greece (GR)	2,279	2,135	- -	1,886	2,370	- -	8,670
Croatia (HR)	- -	- -	- -	1,272	1,407	- -	2,679
Hungary (HU)	1,562	1,414	1,406	1,433	1,447	1,874	9,136
Ireland (IE)	1,866	2,111	1,538	1,462	2,146	2,218	11,341
Iceland (IS)	- -	553	- -	- -	- -	691	1,244
Italy (IT)	1,171	1,487	- -	- -	- -	883	3,541
Lithuania (LT)	- -	- -	- -	- -	1,519	1,938	3,457
Luxembourg (LU)	951	1,043	- -	- -	- -	- -	1,994
Netherlands (NL)	2,167	1,690	1,688	1,572	1,657	1,639	10,413
Norway (NO)	1,881	1,607	1,596	1,394	1,351	1,384	9,213
Poland (PL)	2,027	1,672	1,682	1,576	1,707	1,843	10,507
Portugal (PT)	1,412	1,922	1,995	2,199	1,990	2,002	11,520
Sweden (SE)	1,766	1,745	1,690	1,591	1,300	1,585	9,677
Slovenia (SI)	1,349	1,316	1,338	1,161	1,255	1,127	7,546
Slovakia (SK)	- -	1,388	1,558	1,666	1,727	1,719	8,058
Total	35,707	40,325	34,148	38,361	40,283	39,507	228,331

Appendix 2 Descriptive statistics

	Percent	*N*
Gender		
female	53.5	122,057
male	46.5	106,018
Total	100.0	228,075
Age category		
16-24	14.0	31,738
25-34	14.8	33,624
35-44	17.2	38,979
45-54	17.1	38,847
55-64	16.4	37,206
65-74	12.6	28,714
75+	8.0	18,141
Total	100.0	227,249
Employment status		
self-employed	6.5	14,626
higher service class	6.3	14,185
white-collar workers	20.9	47,052
blue-collar workers	14.9	33,576
unemployed	5.2	11,835
retired	24.5	55,458
in education	8.6	19,496
homemakers	9.1	20,596
disabled	2.3	5,282
other	1.3	2,922
Total	99.6	225,028
Left-right placement		
Left	27.7	63,239
Center	28.82	65,804
Right	30.92	70,598
Missing	12.57	28,690

	Mean	SD	Min	Max	*N*
Economic threat perceptions	5.21	2.39	0	10	217917
Cultural threat perceptions	4.50	2.49	0	10	218073
Education (in years)	12.10	4.03	0	30	225821
Subjective income	2.98	0.86	1	4	222897
Urbanization	3.04	1.21	1	5	227676
Religious involvement	4.40	2.55	0.71	10	227353

methods, data, analyses | Vol. 12(2), 2018, pp. 211-232 DOI: 10.12758/mda.2017.13

Demonstrating How to Best Examine Group-based Segregation: A Statistical and Conceptual Multilevel Approach

Christoph Spörlein[1] *& Elmar Schlueter*[2]

[1] Universität Bamberg
[2] Justus-Liebig-Universität Giessen

Abstract

Segregation between ethnic or sociodemographic groups represents a longstanding key independent and dependent variable for the social sciences. However, researchers have only recently begun to take advantage of inferential rather than descriptive statistical techniques in order to assess various aspects of segregation. Specifically, this paper shows that the multilevel binomial response approach suggested by Leckie et al. (2012) provides a particularly flexible framework for describing and explaining segregation in ways not previously possible. Taking the index of dissimilarity (*D*) as an example we demonstrate how the multilevel binomial response approach helps to reduce the problem of small unit bias, allows to asses segregation at different scales and enables researchers to better understand the role of individual- and contextual-level explanatory variables in shaping segregation. To this end, we employ three case studies focusing on different manifestations of ethnic and gender segregation using survey data from urban, national and cross-national settings.

Keywords: index of dissimilarity, segregation, composition, context, multilevel modelling, simulation

An important question in comparative social science research is this: To what extent and why do members of different groups live or work segregated from another? For example, ethnic residential segregation – broadly defined here as the extent to which members from distinct ethnic groups are unequally distributed across residential areas – is often seen as a core independent variable driving multiple forms of ethnic inequality, e.g. in education or on the labour market (Lieberson 1980). Likewise, several social science approaches seek to understand the factors shaping ethnic residential segregation as dependent variable (Massey 1985, Alba and Logan 1993). Segregation, however, is certainly not limited to occur between members of different ethnic groups or with regard to residential areas only. To name just one further example, a longstanding and influential literature deals with the causes and consequences of differences in the distribution of men and women across occupations and related settings, a phenomenon known as gender segregation in the labour market (Chafetz 1988). Empirically, in order to assess different forms of segregation researchers commonly rely on official census data. For sheer size and scope alone, such data certainly represent a very broad and hence useful empirical source. However, the administrative and financial constraints to obtain census data often still are far from trivial. Also, the availability of census data sometimes is restricted to aggregate data only. While sufficient for several purposes, aggregate data might not always meet the requirements of the research question of interest. At this point, the increasing availability of large-scale survey data in conjunction with recent statistical and computational advances opens up new possibilities for research on segregation. Accordingly, this contribution seeks to illustrate the synergies to be achieved when using publicly available survey data in concert with state-of-the-art inferential methods of data analysis in order to adequately describe and explain segregation in different fields. We do so by demonstrating the virtues of using the multilevel binomial response approach to assess segregation recently developed by Leckie et al. (2012). As we explicate below, this statistical framework enables researchers to draw inferential rather than descriptive conclusions, to account for small unit bias, to assess segregation at multiple scales as well as to evaluate the contribution of explanatory variables at different levels of analysis. Given multiple forms of segregation and researchers' interest to quantify segregation by a single number, today a great variety of different so-called segregation indices is available (Massey and Denton 1998). While we endorse this plurality of segregation measures, for pragmatic reasons here we focus on the index of dissimilarity (*D*) as a particularly well-known and popular measure of segregation.

Direct correspondence to
Christoph Spörlein, Universität Bamberg
E-mail: christoph.spoerlein@uni-bamberg.de

Modelling the Index of Dissimilarity

The index of dissimilarity (D) is perhaps the most widely used measure in the social sciences when interest lies in quantifying the degree to which two groups A and B are unevenly distributed across J units. D often is defined as (Duncan and Duncan 1955)

$$D = \frac{1}{2}\left(\sum_{j=1}^{J}\left|\frac{a_j}{A} - \frac{b_j}{B}\right|\right) \tag{1}$$

Here, a_j is the observed proportion of group A in unit j, b_j the corresponding observed proportion of group B in unit j and A as well as B refer to the total proportions of groups A and B (Duncan and Duncan 1955). D ranges from 0 to 1 where 0 indicates no segregation and 1 describes a scenario with total segregation. Values of D within this range are commonly interpreted as the fraction of either group A or B that would have to change across units J in order to achieve an even distribution across the J units. While intuitively appealing and easy to compute using simple cross-tabulation, researchers long have noticed several limitations of D. For example, researchers typically calculate D from observed proportions. An important drawback of this approach is that it fails to recognize the underlying stochastic processes that generate these proportions (Leckie et al. 2012). This means that even if the allocation of individuals to units (i.e., ethnic minority and majority members to neighborhoods, men and women to occupations) was purely random, D will most likely be non-zero due to random sampling that drives unevenness in the distribution to some non-negligible extent. Further, this upward bias in D is known to systematically vary with the proportions of individuals per unit such that the likelihood of observing highly segregated units is inversely related to unit size (i.e., small cell bias, Carrington and Troske 1997, Allen et al. 2009, Mazza and Punzo 2015). Accordingly, when segregation is investigated for a relatively large number of sparsely populated units, random sampling alone might produce some highly segregated units, which in turn generates a disproportionate upward bias in D. Drawing on earlier work by Goldstein and Noden (2003), Leckie et al. (2012) developed an elegant statistical solution that overcomes these limitations. These authors demonstrate that a binomial response multilevel model effectively takes into account the binomial sampling variation when modelling observed proportions of individual observations in units and reduces the risk of small cell bias. Statistically, this approach takes advantage of multilevel shrinkage (Raudenbush and Bryk 2012) where units with fewer observations contribute less to the estimation of parameters compared to units with more observations. Consider the following basic two-level binomial response multilevel model:

$$
\begin{aligned}
& y_i \sim Binomial\left(n_j, \pi_j\right) \\
& \operatorname{logit}\left(\pi_j\right) = \beta_0 + u_j \\
& u_j \sim N\left(0, \sigma_u^2\right)
\end{aligned} \tag{2}
$$

where y_j denotes the probability that an individual in unit *j* belongs to group A, n_j is the total number of individuals in units *j* and π_j is the unknown underlying proportion of group A in unit *j*. The underlying proportion π_j is determined by $\beta_0 + u_j$ through a logit link. β_0 denotes the intercept and when exponentiated represents the average proportion of group A in the 'median' unit *j*. u_j denote the random effects varying across units *j*. The random effects u_j are central to the multilevel framework of segregation because they effectively serve as a naïve estimator of the degree of segregation across unit *j:* the larger the random effects, the larger the variation of the average proportion of group A across units *j*. Conversely, if u_j is zero, then the proportion of group A across unit *j* is constant and therefore no segregation is observed. Once we obtained the estimates for the model described in equation (2), we can calculate *D* using a simulation approach described in Leckie et al. (2012) to compute adjusted counts per unit where M is the number of iterations. Specifically, the simulation proceeds in four steps that build incrementally:

Step 1: Simulate one value for each of the J unit-level random effects using the model estimate of the unit-level variance $\sigma_u^2 : u_j^{(m)} \sim N\left(0, \sigma_u^2\right)$.

Step 2: Compute the estimated proportion of group A per unit $j : \pi_j^{(m)} : anti - \operatorname{logit}\left(\beta_0 + u_j^{(m)}\right)$.

Step 3: Compute the adjusted counts of group A per unit $j : n_j^{(m)A} = \pi_j^{(\mathrm{m})} n_j$; with the adjusted counts of group B per unit j computed as $n_j - n_j^{(m)A}$.

Step 4: The dissimilarity index is then computed as

$$
D^{(m)} = \frac{1}{2}\left(\sum_{j=1}^{J}\left|\frac{n_j^{(m)A}}{\sum_{j=1}^{J} n_j^{(m)A}} - \frac{n_j^{(m)B}}{\sum_{j=1}^{J} n_j^{(m)B}}\right|\right) \tag{3}
$$

Summarizing the resulting vector of M dissimilarity indices by its mean and the corresponding 95% confidence interval yields the desired measure of uncertainty. In this way, uneveness due to binomial sampling variation respectively small cell bias is adequately taken into account when calculating *D*, with the confidence interval providing additional information about the statistical significance of *D*. However, approaching segregation from a statistical and conceptual multilevel perspective offers additional and equally important advantages. Perhaps most interestingly, the multilevel approach outlined above enables researchers to model segregation as a function of explanatory variables at different levels of analysis. Typical (two-level) applications of multilevel modelling often seek to model between-context variance (e.g., cross-national differences in respondents' average income or explaining school differences in pupils' average math-skills). This level-two variance can potentially be explained by compositional differences across the level-two units, level-two characteristics or a combination thereof (Raudenbush and Byrk 2002, Hox 2010, Snijders and Bosker 2011). Consequently, adding level-one respectively level-two explanatory variables will likely reduce the level-two variance[1]. One issue with this modelling approach lies with the fact that the comparison of nested non-linear models is problematic because the individual level variance is fixed to $\pi^2/3$ (Hox 2010). When including independent variables, parameter estimates of the model will be rescaled in such a way that the variance on the individual level remains constant at $\pi^2/3$. Obviously, this is problematic when these parameter estimates are fundamental to the simulation steps of the multilevel framework. Hence, as one extension of Leckie et al.'s (2012) modelling approach, we aim to remedy this drawback by bringing all models to the same baseline scale of the respective null models through multiplication of all parameter estimates with the squared "scale correction factor" (Hox 2010: 133ff). The scale correction factor is defined as

$$SCF = \sqrt{\frac{\frac{\pi^2}{3} + \sigma^2_{u(0)}}{\frac{\pi^2}{3} + \sigma^2_u + var(\hat{\pi})}} \qquad (4),$$

where $var(\hat{\pi})$ denotes the variance of the linear estimator and the index (0) refers to parameters from the null model (i.e., a model to estimate the unadjusted segregation). This correction is applied throughout all analyses presented in this article.

In terms of multilevel modelling of segregation, a decrease in the random effects means that some part of the observed segregation pattern is due the explanatory variables added to the model[2]. Apparently, this option is particularly advanta-

1 In some instances, adding level-1 variables may increase level-2 variation. Typically, this occurs when variables are added to the model that contain no or only very little between-unit variation (Hox 2010: 74). For instance, the sex distribution across city districts is unlikely to vary substantially thus adding individual's sex may increase the variation on the district level. Dropping variables with little level-2 variation should solve the issue.

2 Kalter (2001) proposed a multinomial logit framework to adjust D for compositional differences across two groups. However, this framework does not take into account small cell bias nor does it enable researchers to add unit-specific explanatory variables of the observed segregation patter ns (e.g., occupational characteristics or neighborhood characteristics).

geous for examining the individual respectively contextual level factors presumed to generate or maintain segregation between groups. At the same time, conceptualizing segregation in a multilevel framework opens up the possibility to model segregation across multiple scales. Thus, in terms of residential segregation, this means that one could model segregation with respect to neighborhoods and cities in one model by introducing a hierarchical city level (level 3) in addition to the neighborhood level (level 2) and individual residents (level 1). Note that this framework also can easily incorporate non-hierarchical segregation structures using a cross-classified design, e.g. occupational and industrial gender segregation (see study 3).

Three Case Studies

In the empirical part of our paper, we present three case studies of modelling *D*. These examples illustrate not only different modelling options offered by the proposed new method, but also provide novel answers to interesting substantial research questions. The first example presented in study 1 uses data from German urban monitoring survey in which German citizens and immigrants were sampled from a large number of city districts. These data enable us to study the extent of ethnic residential segregation between city districts, holding constant socioeconomic differences among respondents and accounting for district-level characteristics. The second example presented in study 2 directs its attention to the field of cross-national research. Using individual data from the European Labour Force Surveys (EU-LFS), we study the degree of ethnic occupational segregation for 15 EU member states that remains after both individual- as well as occupation-level explanatory variables are taken into account. In study 3, the research question of interest for the final example is to determine simultaneously the level of gender occupational and industrial segregation. To this end, we employ a cross-classified multilevel model using a single wave from the German Socio-Economic-Panel Study (GSOEP).

Study 1: Ethnic Residential Segregation

Data and Theory

We study ethic residential segregation using data from the urban monitoring survey program of the city of Duisburg (‘Duisburger Bürgerumfrage’, see GESIS 2017), a large multiethnic city situated in the western part of Germany (see Schlueter, 2011). Focusing on topics such as residents’ satisfaction with the cultural and social infrastructure of the city, these surveys were carried out separately for German citizens and foreigners using random samples of individuals aged 18 years and older selected from the city's population register. For the present purposes and in order to increase sample size, we merged three waves of data spanning the years 2004, 2005 and 2006 (Stadt Duisburg, Amt für Statistik, Stadtforschung und Europaangelegenheiten der Stadt Duisburg, 2007). From the sample of foreigners, we selected only Turkish respondents[3] as they represented the largest ethnic minority group in Duisburg (~24 percent). Our final sample covers 6,352 individuals (level 1), 21 percent of which from Turkish descent, living in one of 46 districts in the city of Duisburg (level 2). The dependent variable in this case study is a dichotomous variable indicating whether respondents are of Turkish descent (1) or of German descent (0).

We employ three theoretical perspectives to describe and explain ethnic residential segregation. Our vantage point is the spatial assimilation model (Massey, 1985), which posits that ethnic minority members are able to convert their socioeconomic resources for renting or acquiring living space that is equally desired by ethnic majority members. According to this approach, the extent of ethic residential segregation should diminish once the socioeconomic resources of group members are taken into account. To this end, we include three individual-level indicators reflecting group compositional differences in socioeconomic resources (highest education attainment [1 = no education to 3 = (Fach-) Hochschulreife], respondent receives unemployment benefits and respondent receives social welfare). For completeness, we also hold constant respondents' age (in years), gender (0 = male, 1 = female), marital status (0 = not married, 1 = married) and household size (number of persons per household). Unlike the spatial assimilation model, the place stratification model holds that ethnic residential segregation centrally is shaped by powerful majority members (e.g. real estate agents, landlords) who seek to constrain ethnic minority members' access to desirable residential spaces (Alba and Logan, 1993). Supposing that a substantial degree of ethnic residential segregation persists even after controlling for differences in the socioeconomic resources of group members, this means that more (less) attractive districts should increase (decrease) ethnic residential segregation. We seek to approximate these assumptions by assessing the desirability of city districts using information on the average living space per person (2005 data) and average rent per square meter (no utilities, 2002 data), presuming that a higher average living space per person respectively higher average rent per square meter indicates more attractive city districts. Further, we take the number of social welfare recipients per 1,000 inhabitants (2005 data, Stadt Duisburg 2007) to indicate less attractive city districts. The third theoretical account we consider is known as the homophily-principle. Shifting attention to group members' ethnic preferences, this approach presumes at its core that ingroup members prefer to dwell among fellow ingroup members (Schelling 1969; McPherson, Smith-Lovin and Cook 2001; Henry, Pralat and Zhang 2011). We approximate

3 Extending this example to multigroup comparison is fairly straightforward using multinomial logistic multilevel models or a series of binomial multilevel models.

this assumption using data on the local ethnic infrastructure represented by the proportion of 'ethnic clubs' per postal code district gathered from the Federal Register for Associations (Justizministerium 2016).

Results

Figure 1 depicts the results for the gross level of ethnic residential segregation and the subsequent adjustments for compositional differences between Germans and Turks as well as contextual differences across city districts. The first two bars of the figure show that the gross level of ethnic residential segregation is fairly similar when calculated based on the standard cross-tabulation approach and the multilevel simulation approach. Both methods result in an index of dissimilarity that approaches a value of 0.40. In addition, the simulation results provide a 95% confidence interval depicted as error bars which range from 0.31 to 0.47. According to the common interpretation of *D*, in order to for the two population groups to be evenly distributed across Duisburg's city districts, roughly 40 percent of the population would need to move between districts. However, adjusting the observed level of residential segregation for potential compositional differences between Germans and Turks in terms of their socioeconomic resources results in a decline to an average of 0.28 (CI=[0.22;0.35]). In other words, around one quarter of the observed level of ethnic residential segregation in Duisburg is accounted for by the average lower socioeconomic positions of Turks relative to Germans – a large effect.

Table 1 presents the results of the multilevel models which provided the parameters for the simulation of the dissimilarity index, specifically, the intercept and the district-level random effect. Assessing the direction of change in segregation after adjustment for compositional differences is easily glimpsed by the reduction of the district-level random effect which decreases from 1.08 to 0.58 (Variance district-level$\times SCF^2 = 0.97 \times 0.60 \sim 0.58$). Hence, even without carrying out the simulation of *D* the change in the district level variance provides an intuitive measure of the change of segregation: the variance on the district level indicates how strongly the average proportion of Turkish residents per city district deviates from the median neighborhood. Thus, a reduction in this variation implies that some fraction of the between-district variation in the proportion of Turkish residents is accounted for differences in the socioeconomic composition of the two groups.

Finally, model 2 incorporates the contextual measures of the local pricing structure and ethnic infrastructure which results in a further decrease in the level of segregation to an average of 0.17 (CI=[0.13;0.21]). Contrary to our expectation, we do not find that the proportion of ethnic clubs is associated with the proportion of Turkish inhabitants per neighborhood. The pricing indicators are more in line with our expectations: city districts with on average larger rooms have lower proportions of Turkish inhabitants whereas the number of social welfare recipients per 1,000 inhabitants is positively associated with a districts' proportion of Turks. Although

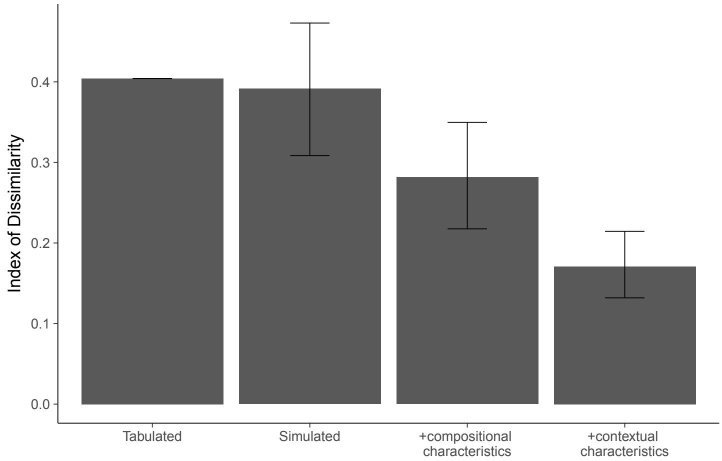


Figure 1 Ethnic residential segregation in Duisburg, calculated based on tabulated data and using multilevel binomial response approach

these associations are present net of individual differences in resource endowment, the associations point towards a primary underlying mechanism, namely that the lower socioeconomic composition of Turks in Duisburg constraints their residential choices which in turn is associated with a large part of the observed segregation patterns. Overall, the adjustment of segregation for compositional and contextual differences reduced the index of dissimilarity by roughly 60 percent[4].

Study 2: Ethnic Occupational Segregation

Data and Theory

In order to study ethnic occupational segregation, we rely on cross-national data from the European Labour Force Survey (LFS) for the EU-15 member states. For this application, we focus on comparing occupational choices of first generation immigrants (i.e., those born outside the respective destination country) to the national population. Specifically, we use data from the 2009 wave covering (self-) employed individuals aged 22 to 57. Occupations are classified according to three-digit ISCO-88 codes which provide a suitable compromise between level of detail (i.e., 131 distinct occupational categories) and individuals per occupational category. Moreover, the analyses will be carried out separately not only by country,

4 Notice that and variance on the neighborhood level is reduced by roughly 80 percent. This difference is due to the non-linear relation between the random effects and the dissimilarity index (Leckie et al. 2012:15).

Table 1 Multilevel modelling of ethnic residential segregation in Duisburg, 2003-2006 (n=6,532)

	M0: gross D		M1: + individual characteristics		M2: + contextual characteristics	
	coef.	s.e.	coef.	s.e.	coef.	s.e.
Intercept	-1.59*	0.16	-2.86*	0.18	-2.78*	0.14
Age			-0.03*	0.00	-0.04*	0.00
Female			-0.43*	0.08	-0.44*	0.08
Married			1.57*	0.11	1.56*	0.11
Household size			0.50*	0.03	0.50*	0.03
Group compositional differences						
Education						
No education			2.40*	0.14	2.38*	0.14
Hauptschule			0.70*	0.12	0.68*	0.12
Realschule (ref.)						
(Fach-)Hochschulreife			0.03	0.15	0.04	0.15
Receives unemployment benefits			0.99*	0.13	0.97*	0.13
Receives social welfare			0.72*	0.19	0.71*	0.19
Contextual differences						
Average room size					-0.08*	0.03
Average price per qm					-0.41	0.23
Social welfare recipients per 1,000 inhabitants					0.01*	0.00
Proportion of ethnic clubs					-0.94	1.29
Variance neighborhood level	1.08		0.97		0.38	
SCF^2	-		0.60		0.53	
R^2 neighborhood level	-		0.55		0.84	

Note. All variables (with the exception of “education”) grand-mean centered. Comparisons across models require multiplication of M1 and M2 coefficients with the squared scale correction factors.

but also by gender – an important category in research on labour market segregation. Our final sample includes 1,082,025 individuals (11.2 percent of which are immigrants) living in one of the EU-15 member states. The dependent variable in this case study is a dichotomous variable indicating whether respondents were born outside their country of residence (1) or born in the country of residence (0).

Labour market outcomes such as occupational sorting typically result from matching processes between employers wanting to fill vacancies with suitable candidates and employees expecting to receive adequate compensation for the skills they offer (Kalleberg and Sørensen, 1979). Systematic differences in occupational sorting between immigrants and the majority population may therefore result from (1) between-group differences in the skills they offer or (2) preferences of employers for individual characteristics that go beyond skill endowment (i.e., discrimination; Granato and Kalter 2001). Since discriminatory explanations are notoriously difficult to uncover with large-scale survey data, we focus on the first aspect, namely compositional differences between immigrants and the majority population in terms of relevant skills. Central to group differences regarding skills will be educational attainment as a first crude approximation where higher levels of education are assumed to be associated with higher skill levels. This approximation obviously ignores substantial variation in labour market skills within educational categories. We try to improve the approximation by including occupational characteristics that are correlated with differences in skill level. For example, two occupations may be chosen by individuals with similar educational attainment profiles. But these occupations differ along other dimensions (e.g., the prevalence of temporary employment contracts) that make them more or less attractive to the higher skilled employees and thereby help in explaining group differences in occupational sorting beyond mere compositional differences in the absence of adequate data. Hence, when trying to account for the observed degree of ethnic occupational segregation, we include the following individual characteristics (i.e., compositional differences between immigrants and the majority population, level 1) as well as contextual characteristics (i.e., differences between occupational categories, level 2). For the first set of characteristics, measures of age (in years), marital status (0=not married, 1=married), nationality (0=nationalized, 1=non-national), educational attainment (0=ISCED to 6=ISCED 6), weekly work hours and full-time employment (0=part-time, 1=full-time). In contrast to the data used in case study 3, the EULFS includes few relevant occupational characteristics. We therefore rely on aggregating country-specific individual characteristics for each occupational category: the percentage of firms employing more than 10 workers, the percentage of workers holding temporary contracts and the percentage employed in non-shiftwork. Results for the simulated index of dissimilarity D are based on gender- and country-specific multilevel binomial response models where employees (level 1) are hierarchically nested in 131 occupational categories (level 2).

Results

Figure 2 visualizes the results for modelling ethnic occupational segregation separately for men (upper panel) and women (lower panel) across 15 European-Union countries. To begin with, we note that the figure shows substantial cross-national variation in the extent of D. For males, the results for simulating D from the initial models without individual- respectively occupational-level explanatory characteristics range from a minimum of 0.18 (Belgium) to a maximum of 0.49 (Greece). For females, the minimum in ethnic occupational segregation equals 0.15 (Belgium), while its maximum is 0.52 (Greece). To illustrate, these numbers could be taken to imply that in Belgium, 18% of the first generation male immigrants and majority members would need to change between occupational categories in order to achieve an equal distribution across all occupations. However, the results from the subsequent models demonstrate that the extent of ethnic occupational segregation is substantially reduced once the previously discussed explanatory variables are taken into account. For all countries and for both males and females, controlling for compositional characteristics of the individual employees uniformly results in a decrease of *D*. For example, the largest drops in *D* are found for Italy (for male employees, $\Delta D = 0.24$; for female employees, $\Delta D = 0.19$). To reiterate the logic of the underlying modelling approach, we note that parts of the level-2 variance, which in this case reflects how strongly the proportion of immigrants varies across occupations, are accounted for by differences in, for instance, educational profiles or weekly work hours between immigrants and the respective host society populations. Conversely, the remaining level-2 variance suggest that even after accounting for these compositional differences, immigrants are still disproportionately more often working in some occupations rather than others. This implies that there are either compositional differences we haven't picked up yet and/or that these differences can be explained by systematic difference of occupations themselves. The figure also shows that adding the occupation-level characteristics to the model leads for many countries to a further, albeit relatively small decrease in the extent of ethnic occupational segregation. Interestingly, between countries the data reveal a heterogeneous pattern of results between occupational characteristics and the proportion of (male) immigrants in each country-specific occupational category. For example, whereas an occupations higher prevalence of shift work is positively associated with a higher share of immigrants in Germany, Greece, Italy, Spain and the UK, no such correlation is present in the remaining EU-15 countries. Similarly, in some countries immigrants seem to overrepresented in occupations typically present in smaller firms in some countries: in Belgium, Greece, Italy and Spain, the results show that the higher an occupation's proportion of individuals working in firms with more than 10 employees, the smaller that occupation's proportion of immigrants. However, in the remaining countries this association is virtually absent[5]. Collectively, these results could be taken to explore potential country-level moderators of the divergent relations between the predictors and the proportion of (male) immigrants in the occupational categories.

5 Table A1 in the appendix summarizes the pattern of results. Detailed results are available upon request.

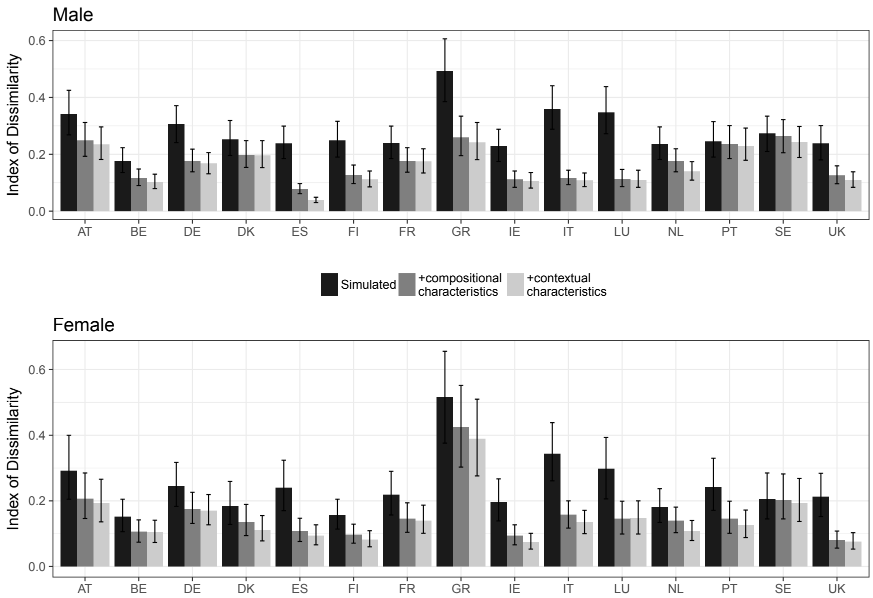


Figure 2 Ethnic occupational segregation in EU-15 countries based on multi-level simulation approach

Study 3: Occupational and Industrial Gender Segregation

Data and Theory

For the third case study, we rely on data from the German Socio-Economic Panel Study (GSOEP, Wagner et al. 2007), which has been collected annually since 1984 as a probability-based sample of households. We use the 2011 wave and restrict the sample to (self-) employed individuals aged 25 to 54. Information on occupations is again based on three-digit ISCO-88 codes. Because we are also interested in estimating the level of gender industry segregation that is independent of occupational segregation, we rely on the division categories of the NACE classification of industries which comprises a total of 62 categories (e.g., 'crop and animal production', 'manufacturing of electrical equipment' or 'education'). In total, the sample covers 7,802 employees working in 108 occupations and 62 industries. The dependent variable in this case study is a dichotomous variable indicating whether respondents were female (1) or male (0).

Similar to the mechanisms that generate patterns of ethnic occupational segregation, occupational segregation is a result of women and men systematically sorting into different occupations. The reasons for this differential sorting are broadly associated with gender-specific preferences in occupational characteristics as well as differences in (anticipated) life course pressures (Ostner 1990; Hakim 2002; Padavic and Reskin 2002). According to socialization theory, occupational preferences are established in earliest socialization with individuals adopting gendered skills to varying degrees. Gendered preferences may lead individuals to opt for jobs where these skills are more advantageous such as occupations with a strong "social" or "caring" component in the case of women; occupations which are typically part of the service sector (Busch 2013). In addition, different stages in the life course are associated with specific pressure on individuals to reconcile family and employment (Filer 1989; Tam 1997). These pressures are especially marked for women with (small) children who therefore more often work part-time or in jobs with higher flexibility (Glass and Camarigg 1992; Bush 2013; Cha 2013).

Hence, in order to account for the extent of occupational and industry segregation in Germany, we include measures that aim to capture differences in life-course stages and occupational characteristics indicative of job higher flexibility. In terms of life-course stage, we include individual-level measures for marital status (1=married, 0=else), household type (1= single household, 2= single parent household, 3= two person household, no children present, 4= two parent household, at least one children younger than 16 years present, 5= two parent household, children 16 or older present, 6=other), the number of children present in the household who are younger than 16, the total years of full-time work experience and the number of years individuals worked with their current employer. Flexibility differences are captured using the following individual-level characteristics: respondent works part-time, respondent holds a managerial position and works in service industry. In addition, we include occupation-level characteristics which were computed from the SOEP data: the percentage of public employees, percentage working in the service industry, percentage of individuals working in the occupation they trained for, average status of occupation (based on ISEI scores), average company size and average job autonomy. And finally, we also include respondent's education based on the six category ISCED 1997 classification. Notice that occupations and industries are not necessarily nested within another; for example, a white or blue collar workers can certainly work in different industries, and vice versa. Thus, a more realistic view is to consider employees to be situated in a cross-classification of jobs and industries, and this is why we use a non-hierarchically cross-classified multilevel model (Raudenbush and Bryk 2012). Accordingly, in this example, we take employees (level 1) to be non-hierarchically nested in both occupations (level 2a) and industries (level 2b). Our results are based on cross-classified multilevel binomial response models where employees are non-hierarchically nested in 108 occupations and 62 industries.

Results

The main results for this case study are presented in Figure 3. The index of dissimilarity based on cross-tabulated data is calculated as 0.52 for occupational gender segregation and 0.39 for industrial gender segregation. The corresponding values from the multilevel simulation approach are 0.45 (CI=[0.38;0.52]) for occupational and 0.18 (CI=[0.13;0.25]) for industrial gender segregation. Hence, there is considerably less agreement in the extent of segregation compared to the findings presented for residential segregation above. That is because the calculation based on cross-tabulated data is only two-dimensional and thus cannot take into account deviations from unevenness due to some other but possibly related grouping factor. The same is not true for multilevel simulation approach: here, additional grouping factors are taken into account by simply modelling them. The corresponding random effect is estimated net of other random effects present in the models. The differences between the tabulated and simulated indices of dissimilarity thus indicate that some part of occupational segregation is due to industrial segregation and vice versa. Though, apparently it is primarily industrial gender segregation that is artificially inflated due to not taking into account occupation-level random effects.

The following bars in Figure 3 depict the simulated dissimilarity index when adding employee characteristics to the models (see Table 2, model M1 for detailed results). As expected, differences in employee characteristics explain parts the variance in

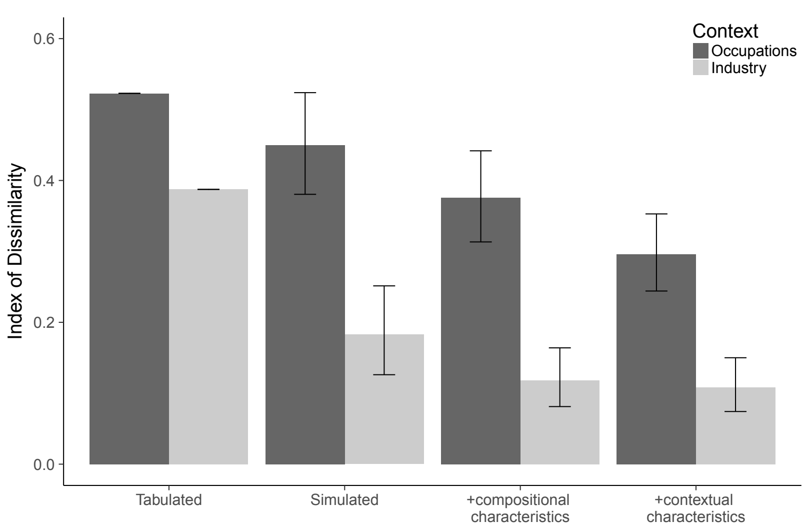


Figure 3 Occupational and industrial gender segregation in Germany, calculated based on tabulated data and using cross-classified multilevel binomial response models

the proportion of women across occupations and industries: levels of segregation decrease to 0.38 (CI=[0.31;0.44]) for occupations and to 0.12 (CI=[0.08;0.16]) for industries. Thus, on average 16 percent ([0.45-0.38]/0.45) of occupational gender segregation are due to employees with specific characteristics differentially sorting across occupations: for example, some occupations are more frequently composed of individuals working part-time or in service industries. And because these characteristics are more prevalent among women, the inclusion of their attributes in the simulation models accounts for some of the observed unevenness in the gender distribution across occupations. Similarly, differences in employee composition account for roughly 33 percent of industrial gender segregation. And finally, adding characteristics of occupations to the model further decreases the simulated segregation to an average of 0.30 (CI=[0.24; 0.35]) for occupations and to 0.11 (CI=[0.07; 0.15]) for industries. According to the estimates presented in M2, occupations with a higher percentage of employees working in service industries also tend to have a higher proportion of women working in them. None of the other occupational characteristics covary with the proportion of women per occupation. Occupational characteristics account for an additional 15 percentage points in occupational gender segregation and another 5 percentage points in industrial gender segregation through differences across industries regarding their occupational make-up. Even though we included a broad range of factors associated with differences in life-course stages and flexibility demands, especially the extent of occupational gender segregation remaining is substantial: around one third of female or male employees would need to change occupations to arrive at an even distribution.

Table 2 Multilevel modelling of occupational and industry gender segregation, GSOEP 2011 (n=7,802)

	M0: gross D		M1: + individual characteristics		M2: + contextual characteristics	
	coef.	s.e.	coef.	s.e.	coef.	s.e.
Intercept	-0.50*	0.17	-0.10	0.29	0.01	0.28
Educational attainment			-0.01	0.03	-0.01	0.03
Group compositional differences						
Married			0.48*	0.08	0.48*	0.08
Household type (ref. other)						
Single household			-0.10	0.26	-0.09	0.26
Single parent household			1.28*	0.29	1.30*	0.29
Two person household			0.14	0.26	0.15	0.26
Two parent household, at least one child younger than 16 present			-0.71*	0.27	-0.70*	0.30
Two parent household, children 16 or older present			-0.24	0.26	-0.23	0.26
Number of children younger than 16			-0.17*	0.07	-0.17*	0.07
Total years of full-time work experience			-0.07*	0.01	-0.07*	0.01
Number of years worked with current employer			0.02*	0.01	0.02*	0.01
Works part-time			2.20*	0.11	2.19*	0.11
Holds managerial position			-0.50*	0.09	-0.51*	0.09
Works in service industry			0.46*	0.14	0.35*	0.14
Contextual differences						
Percentage of public employees					-0.75	0.63
Percentage working in service industry					2.52*	0.48
Percentage working in occupation they trained for					-0.27	0.64
Average occupational ISEI					0.01	0.01
Average job autonomy					-0.12	0.30
Variance occupation level	2.03		1.67		1.07	
Variance industry level	0.24		0.13		0.13	
SCF^2	-		0.74		0.63	
R^2 occupation level	-		0.39		0.67	
R^2 industry level	-		0.60		0.66	

Note. All variables grand-mean centered. Comparisons across models require multiplication of M1 and M2 coefficients with the squared scale correction factors.

Discussion

In this article we sought to demonstrate how using a standard multilevel binomial response model in an atypical way enables researchers to overcome several limitations that long have hindered research on segregation. In following Leckie et al. (2012), we showed how the upper-level variances from a binomial multilevel model can be effectively used as accurate measure of ethnic and gender segregation. Further, by employing simulation techniques we then converted this measure into the popular and well-known index of dissimilarity D. This methodological strategy helped not only to avoid the common inflation of D due to small unit bias. In addition, the novel approach also enabled us to assess segregation simultaneously at different scales and to examine the contribution of explanatory variables at multiple, statistically appropriate levels of analysis[6]. Although our primary focus in this paper was methodological, our illustrative case studies yielded several substantial findings that deserve enhanced attention in subsequent research. Specifically, to the literature on ethnic residential segregation this study adds the insight that controlling for individual-level differences in group members' socioeconomic resources drastically reduces the degree of residential segregation (Teltemann, Dabrowski and Windzio, 2015). Unlike previous research, our results show that even after an array of individual-level differences is taken into account, contextual-level characteristics still make a significant contribution to ethnic residential segregation. Relatedly, this study also extends previous knowledge on ethnic occupational segregation. Our findings not only show a substantial decrease in ethnic occupational segregation across several countries once various individual-level factor are taken into account. In addition, the results also offer new insights of the role occupation-level characteristics play in shaping differences between ethnic majority- and minority at the labour market. With regard to the literature on gender segregation at the labour market, this article is the first that assesses segregation simultaneously at different levels using a cross-classified multilevel model. Doing so yielded the novel finding that what appears at first sight as different distribution of men and women across occupations should be better understood as simultaneous segregation not only at the level of occupations, but also at the level of industries. Apart from these substantial contributions, future methodological developments might expand our knowledge in several directions. For example, in focusing on D, we have explored the issue of using a multilevel inferential framework for a single index of segregation only. However, it is well-known that research on seg-

---

6 We refer readers interested in applying the methods described here for their own needs to Spörlein, C. (2016). multi.correct: An R package to calculate and correct the Index of Dissimilarity using multilevel/random effects models, available at https://github.com/chspoerlein/multi.correct.git . Simply type multi_correct after loading the package to inspect the code or ?multi_correct for the help file and code examples.

regation offers a particularly rich array of different segregation measures (Massey and Denton, 1985). Indeed, the statistical approach applied in this paper appears to be suitable to several alternative measures of segregation, such as the prominent isolation- respectively interaction-index (see Leckie et al., 2012) or the index of net difference (ND, see Lieberson 1969). It also appears promising to apply the present approach for assessing segregation phenomenon between multiple ethnic or demographic groups. For ease of exposition, in this contribution we restricted our focus on modelling segregation for two groups only. Yet by extending the binomial to a multinomial response model the present approach is also capable to provide accurate estimates of segregation between multiple groups (Jones et al. 2015, Reardon and Firebaugh 2002). Collectively, the insights resulting from such methodological developments will help to better inform our theoretical understanding of the extent and the sources underlying social segregation.

References

Allen, R., Burgess, S., & Windmeijer, F. (2009). More Reliable Inference for Segregation Indices. *Technical Report 216. The Centre for Market and Public Organisation*. University of Bristol.

Amt für Statistik, Stadtforschung und Europaangelegenheiten der Stadt Duisburg (2007). *Duisburg Citizen Survey*. GESIS Data Archive: ZA4348, ZA4349, ZA4515, ZA4516, ZA4517, ZA4522, ZA4523, ZA4524.

Busch, A. (2013). *Die berufliche Geschlechtersegregation in Deutschland. Ursachen, Reproduktion, Folgen*. Wiesbaden: Springer VS.

Carrington, W. J., & Troske, K.R. (1997). On Measuring Segregation in Samples with Small Units. *Journal of Business & Economic Statistics* 15, 402-409.

Cha, Y. (2013). Overwork and the Persistence of Gender Segregation in Occupations. *Gender & Society* 27, 158-184.

Chafetz, J.S. (1988). The Gender Division of Labor and the Reproduction of Female Disadvantage: Toward an Integrated Theory. Journal of Family Issues 9, 108–131.

Deutsch, J., & Silber, J. (2005). Comparing Segregation by Gender in the Labor Force across Ten European Countries in the 1990s: An Analysis based on the Use of Normative Segregation Indices. *International Journal of Manpower* 26, 237-264.

Duncan, O.D., & Duncan, B. (1955). A Methodological Analysis of Segregation Indexes. *American Sociological Review* 20, 210-217.

Filer, R. (1989). *Occupational Segregation: Compensating Differentials, and Comparable Equity: Empirical Inquiries*. Washington, D.C.: National Academy of Science.

Glass, J., & Camarigg, V. (1992). Gender, Parenthood, and Job-Family Compatibility. *American Journal of Sociology* 98, 131-151.

Glitz, A. (2014). Ethnic Segregation in Germany. *Labour Economics* 29, 28-40.

Goldstein, H., & Noden, P. (2003). Modelling social segregation. *Oxford Review of Education* 29, 225–237.

Granato, N., & Kalter, F. (2001). Die Persistenz ethnischer Ungleichheit auf dem deutschen Arbeitsmarkt: Diskriminierung oder Unterinvestition in Humankapital? *Kölner Zeitschrift für Soziologie und Sozialpsychologie* 53, 497-520.

Hakim, C. (2002). Lifestyle Preferences as Determinants of Women's Differentiated Labor Market Careers. *Work and Occupations* 29, 428-459.

Henry, A. D., Pralat, P., & Zhang, C. (2011). Emergence of segregation in evolving social networks. *Proceedings of the National Academy of Sciences of the United States of America* 108, 8605-8610.

Hox. J. J. (2010). *Multilevel Analysis: Techniques and Applications*. New York: Routledge.

Iceland, J., & Nelson, K.A. (2008). Hispanic Segregation in Metropolitan America: Exploring the Multiple Forms of Spatial Assimilation. *American Sociological Review* 73, 741-765.

Iceland, J., & Sharp, G. (2013). White Residential Segregation in U.S. Metropolitan Areas: Conceptual Issues, Patterns, and Trends from the U.S. Census, 1980 to 2010. *Population Research and Policy Review* 32, 663-686.

Justizministerium. (2016). Common register portal of the German federal states. https://www.handelsregister.de/rp_web/mask.do?Typ=e. Retrieved: 22.12.2016.

Kalter, F. (2001). Die Kontrolle von Drittvariablen bei der Messung von Segregation: Ein Vorschlag am Beispiel der familialen Assimilation von Migranten. *Zeitschrift für Soziologie* 30, 452-464.

Karmel, T., & MacLachlan, M. (1988). Occupational Sex Segregation – Increasing or Decreasing? *Economic Record* 64, 187-195.

Kohn, M.L. (1987). Cross-national research as an analytic strategy: American Sociological Association, 1987 Presidential Address. *American Sociological Review* 52:713-731.

Leckie, G., Pillinger, R., Jones, K., & Goldstein, H. (2012). Multilevel Modeling of Social Segregation. *Journal of Educational and Behavioral Statistics* 37, 3-29.

Massey, D.S., & Denton, N.A. (1987). Trends in the Residential Segregation of Blacks, Hispanics, and Asians: 1970-1980. *American Sociological Review* 52, 802-825.

Massey, D.S., & Denton, N.A. (1988). The Dimensions of Residential Segregation. *Social Forces* 67, 281-315.

Massey, D.S., White, M.J., & Phua,V.-C. (1996). The Dimensions of Segregation Revisited. *Sociological Methods and Research* 25, 172-206.

Mazza, A., & Punzo, A. (2015). On the Upward Bias of the Dissimilarity Index and Its Corrections. *Sociological Methods and Research* 44, 80-107.

McPherson, M.,Smith-Lovin, L., & Cook, J.M. (2001). Birds of a Feather: Homophily in Social Networks. *Annual Review of Sociology* 27, 415-444.

Ostner, I. (1990). Das Konzept des weiblichen Arbeitsvermögens. In Arbeitskreis Sozialwissenschaftliche Arbeitsmarktforschung (Ed). *Erklärungsansätze zur geschlechtsspezifischen Strukturierung des Arbeitsmarkts* (p. 22-39). Paderborn: SAMF.

Padavic, I., & Reskin, B.F. (2002). *Women and men at work*. 2nd ed. Thousand Oaks et al.: Pine Forge Press.

Raudenbush, S.W., & Byrk, A.S. (2002). *Hierarchical Linear Models: Applications and Data Analysis Methods*. Thousand Oaks: Sage.

Reardon, S.F., & Firebaugh, G. (2002). Measures of Multigroup Segregation. *Sociological Methodology* 32, 33-67.

Schelling, T.C. (1969). Models of Segregation. *The American Economic Review* 59, 488-493.

Snijders, T.A.B., & Bosker, R. (2011). *Multilevel Analysis: An Introduction to Basic and Advanced Multilevel Modeling*. Thousand Oaks: Sage.

Stadt Dusiburg. (2007). *Sozialbericht 2007*. Duisburg: Amt für Soziales und Wohnen.

Teltemann, J., Dabrowski, S., & Windzio, M. (2015). Räumliche Segregation von Familien mit Migrationshintergrund in deutschen Großstädten: Wie stark wirkt die sozioökonomische Restriktion? *Kölner Zeitschrift für Soziologie und Sozialpsychologie* 67, 83-103.

Tam, T. (1997). Sex segregation and occupational gender inequality in the United States: Devaluation or specialized training? *The American Journal of Sociology* 102, 1652-1692.

Appendix

Table A1: Ethnic occupational segregation in EU-15 countries based on multi-level simulation approach. The table shows significantly ($p < .05$) positive or negative associations of the predictors with the proportion of immigrants per occupational categories

	Age	Non-national	Education	Married	Employed fulltime	Work hours	Average company size	Proportion holding temporary contracts	Proportion without shift work
AT	-	+	-	+		-		+	
BE	-	+	-	+	-		-	+	
DE	-	+	-	+		-			-
DK	-	+	+	+					
ES		+	+	+			-	+	-
FI	-	+		+					
FR	+	+		+					
GR	-	+	+	+			-	-	-
IE	+	+	+	+	-	-			
IT	-	+	+	+			-		-
LU	+	+		+					
NL		+	-	+				+	
PT	-	+	+	+	-				
SE	-	+	+	+				+	
UK	-	+	+	+	-	-			-

methods, data, analyses | Vol. 12(2), 2018, pp. 233-264 DOI: 10.12758/mda.2017.05

Surpassing Simple Aggregation: Advanced Strategies for Analyzing Contextual-Level Outcomes in Multilevel Models

Dominik Becker[1], *Wiebke Breustedt*[2] & *Christina Isabel Zuber*[3]

[1] *University of Tuebingen*
[2] *University of Duisburg-Essen/University of Cologne*
[3] *University of Konstanz*

Abstract

This article introduces two advanced analytical strategies for analyzing contextual-level outcomes in multilevel models: the multilevel SEM and the two-step approach. Since these strategies are seldom used in comparative survey research, we first discuss their methodological and statistical advantages over the more commonly applied approach of group mean aggregation. We then illustrate these advantages in an empirical analysis of the effect of citizens' support for democratic values at the individual level on a contextual-level outcome – the persistence of democracy – drawing on data from the World Values Survey and the Quality of Government project. Whereas we found no significant effect of support for democratic values in the model using simple group mean aggregation, citizens' support for democratic values was a significant predictor of democracies' estimated survival rate when applying latent aggregation in multilevel SEM and the two-step approach. The article corroborates previous concerns with simple aggregation and demonstrates how researchers can improve the validity of their analyses of contextual-level outcomes by using alternative strategies of aggregation.

Keywords: transformational mechanisms, contextual level outcomes, multilevel analysis, sampling error, democratic stability, democratic values

Despite significant methodological advancements, comparative social scientists continue to face the question of how to adequately test theoretical multilevel models empirically. Hierarchical modeling has evolved into a canonical statistical technique for regressing an individual-level variable on individual- and contextual-level predictors. There is no agreement when it comes to multilevel models where the dependent variable is analytically located on the contextual level, though.

Many comparative studies 'solve' this problem through measures of central tendency – such as the average – or the distribution of the data – such as percentages. They then use these aggregates as predictors for the contextual-level dependent variable (for examples, see Fails & Pierce, 2010; Lim, Bond, & Bond, 2005; Muller & Seligson, 1994). This approach has been criticized on both statistical and methodological grounds. Croon and van Veldhoven (2007) demonstrated that group mean aggregation may lead to biased estimates. Griffin (1997) argued that the aggregation procedure needs to take into account the complex theoretical relationships of independent variables at different levels of analysis. When applying simple aggregation, researchers may run the risk of drawing invalid conclusions about how individual-level predictors affect contextual-level outcomes (Snijders & Bosker, 1999).

Given these criticisms, researchers have proposed two more advanced strategies for analyzing contextual-level outcomes in multilevel models: the multilevel SEM and the two-step approach. Since multilevel SEM and the two-step approach are seldom used in comparative survey research, the article seeks to motivate researchers to improve the validity of their inferences when analyzing contextual-level outcomes by going beyond simple aggregation. In the following section, we introduce the methodological and statistical advantages of these two alternative techniques over the group means approach. In our analysis, we illustrate these advantages in an empirical study of the effect of citizens' support for democratic values at the individual level on a contextual outcome – the persistence

Acknowledgments

This paper benefited from feedback on several occasions including the SOCLIFE Graduate School at the University of Cologne, the annual conference of the European Survey Research Association (ESRA) in Lausanne and the joint conference of the International Political Science Association (IPSA) and the International Association for Political Science Students (IAPSS). In addition, we would like to thank Andreas Hartung and Steffen Hillmert for their helpful comments on our event-history models. Finally, we would like to thank two anonymous reviewers whose valuable suggestions helped to improve the overall quality and consistency of our manuscript.

Direct correspondence to

Dominik Becker, University of Tuebingen, Institute of Sociology, Wilhelmstr. 36, 72074 Tuebingen, Germany

E-mail: dominik.becker@ifsoz.uni-tuebingen.de

of democracy. We draw on data from the World Values Survey and the Quality of Government project and study 98 countries between 1946 and 2014. We compare the regression coefficients and confidence intervals of our individual-level predictor – support for democratic values – on democracies' persistence when applying the three methods. Whereas we found no significant effect of support for democratic values in the model using simple group mean aggregation, citizens' support for democratic values was a significant predictor of democracies' estimated survival rate when applying multilevel SEM and the two-step approach. In the final section we therefore conclude that comparative researchers who use simple group mean aggregation when regressing a contextual outcome on individual level predictors may run the risk of wrongly rejecting their hypothesis of interest.

Methodological Foundation and Statistical Background

Testing theoretical multilevel models with contextual-level outcomes poses two challenges. From a methodological point of view, researchers need to establish close correspondence between the theoretical multilevel mechanism and its empirical measurement. From a statistical perspective, they need to choose a method that is both valid and reliable for aggregating the individual-level predictors. In the following, we discuss the methodological foundations of multilevel analysis of macro-level social phenomena. We then proceed to introduce and compare three analytical strategies for analyzing contextual level outcomes: simple manifest group mean aggregation, latent aggregation through multilevel SEM, and the two-step approach. The results of the comparison are summarized in Table 1 at the end of this chapter.

Methodological Foundation

According to the paradigm of structural individualism (Udehn, 2002), the ultimate goal of the social sciences is to explain social phenomena on the contextual – or macro – level as a consequence of individuals' social actions on the individual – or micro – level. Structural individualism distinguishes three explanatory mechanisms (see Figure 1) (Hedström & Swedberg, 1998; Tranow, Beckers, & Becker, 2016). Situational mechanisms (1) link the objective characteristics of the social situation to the subjective expectations and evaluations of individuals. Action-formation mechanisms (2) explain individuals' actions given their subjective definition of the situation. This is a pure micro-level explanatory step. Transformational mechanisms (3) reconstruct how individuals' actions aggregate to create a new social situation. They thereby re-link the micro level to the macro level.

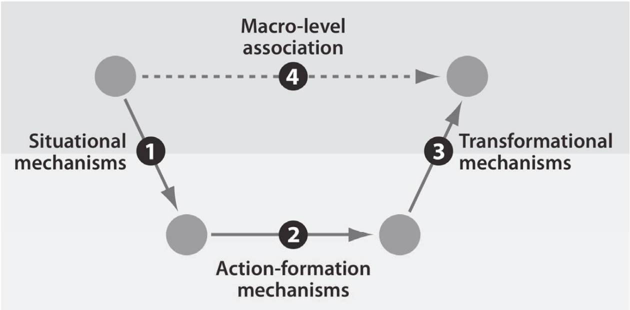


Figure 1 The social mechanisms of social science explanations.
Source: Hedström & Ylikoski (2010, p. 53)

Studying these theoretical mechanisms empirically is not straightforward. Multilevel modeling (Bryk & Raudenbush, 1992; Hox, 2010) is a well-established statistical tool for testing situational and action formation mechanisms, that is, explanations that link social situations to individuals' expectations, evaluations, and actual decisions (Becker, Beckers, Franzmann, & Hagenah, 2016). By contrast, micro-to-macro (or, more technically, level-one to level-two) explanations constitute a blind spot of conventional multilevel analysis (henceforth MLA)[1] as transformational mechanisms are more difficult to analyze empirically (Opp, 2011; Raub, Buskens, & van Assen, 2011).

Three Analytical Strategies

The simple group means approach

When studying multilevel models with contextual-level outcomes, a common approach (Lim et al., 2005) is to aggregate all level-one variables (hereafter L1) to level-two variables (hereafter L2) by computing their group-specific arithmetic means. This manifest aggregation is followed by an L2-only regression (see Figure 2).

Methodologically, this method models neither situational nor action-formation mechanisms and accounts for transformational mechanisms via (manifest) aggregation (see Figure 2). Statistically, Croon and van Veldhoven (2007) have shown that this procedure only yields valid estimates if the L1 variance of the aggregated variables is zero. If the L1 variance is larger than zero, simple group mean aggregation yields biased estimates. In cross-national comparative survey research, this

1 In accordance with previous research, we use the terms 'conventional' or 'standard' multilevel analysis to describe hierarchical modeling techniques that are restricted to the analysis of level-one outcomes (Bennink, Croon, & Vermunt, 2013, 2015; Lüdtke et al., 2008; Preacher, Zyphur, & Zhang, 2010).

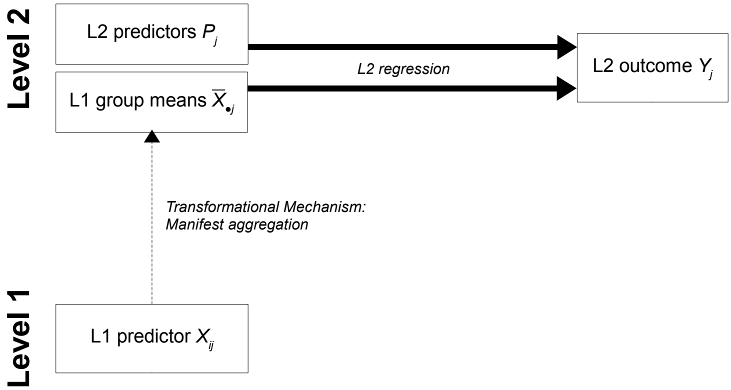


Figure 2 The simple group means approach

is often the case because individuals are sampled from a finite population and a specific constellation of individuals is selected to measure the L2 construct (Lüdtke et al., 2008). Since manifest aggregation does not control for these sampling errors, the observed group average (measured, for instance, in terms of group-specific arithmetic means) may be an unreliable measure of the unobserved true group average. In addition, the observed group average completely obscures the heterogeneity within groups. Therefore, if effects of observed group averages on L2 outcomes are of interest, estimates of both these effects and of other L2 predictors are likely to be biased when applying the simple group means approach (Bennink et al., 2013, 2015; Shin & Raudenbush, 2010).

The multilevel SEM approach

Multilevel SEM avoids these statistical problems by replacing manifest with latent aggregation (see Figure 3). Assume that we observe a manifest L1 variable X_{ij} for individuals i in countries j. X_{ij} is used to predict a manifest L2 outcome Y_j along with other L2 predictors P_j. Following the simple group means approach, X_{ij} is aggregated from L1 to L2 by computing group-specific arithmetic means $\overline{X_{\bullet j}}$, which are not corrected for sampling error. In a second step, $\overline{X_{\bullet j}}$ are used to predict Y_j controlled for P_j (adapted from Marsh et al., 2009):[2]

$$Y_j = \beta_0 + \beta_1 \overline{X_{\bullet j}} + \beta_2 P_j + u_{0j} \tag{1}$$

2 The notation by Marsh et al. (2009) implies group mean centering of all L1 predictors to account for a reference-group effect (in their example, this is the dependence of student academic self-concept on class-average achievement). Since our substantive application does not include a reference-group effect, we present the general notation without group mean centering. In addition, we use standard multilevel notation for the L2 residual variance.

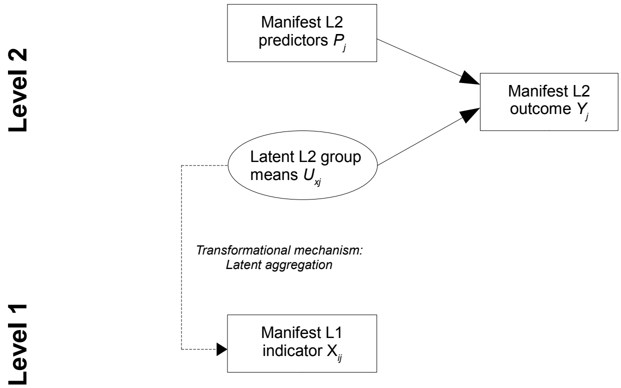


Figure 3 Latent aggregation in multilevel structural equation modeling

By contrast, multilevel SEM regards the actual group mean on L2 as an unobserved latent variable U_{xj} (which must not be confused with L2 residual error u_{oj}) that can only be estimated with error by the L1 indicators (Marsh et al., 2009). Following the conventions of SEM, the L2 latent means of the L1 observations are therefore depicted by ovals in Figure 3. While the simple group means approach treats the L2 group mean as a simple composite or index score of the L1 observations, multilevel SEM assumes the unobserved latent group means to *cause* the observed L1 values (Lüdtke et al., 2008).[3]

Multilevel SEM proceeds in two steps: First, an L2 latent variable U_{xj} is estimated. It is assumed to be the cause of X_{ij} at L1. In a second step, U_{xj} is used to predict the L2 outcome Y_j along with the other L2 predictors P_j:[4]

$$Y_j = \beta_0 + \beta_1 U_{xj} + \beta_2 P_j + u_{0j} \tag{2}$$

The aggregated L2 construct is a measure of the unobserved true group mean. Its reliability is a function of the relative share of the L2 variance weighted by the group-specific number of observations (Lüdtke et al., 2008):

$$\frac{\tau_x^2}{\tau_x^2 + \left(\sigma_x^2 / n_j\right)} \tag{3}$$

3 This points to the difference between formative and reflective models in measurement theory. Whereas formative latent variable models are already established in single-level measurement models (Diamantopoulos & Winklhofer, 2001), it remains unresolved whether formative latent aggregation is equally possible.

4 Additional controls for measurement error can be integrated easily (Marsh et al., 2009). For the sake of simplicity, our analysis of democratic persistence is limited to latent aggregation without controlling for measurement error.

As in conventional hierarchical modeling, σ^2_x denotes the L1 part and τ^2_x the L2 part of the variation of the respective indicator(s), whereas n_j refers to the group-specific number of observations.

By estimating a latent L2 variable U_{xj} as in (2), the variance of the L1 indicator is partitioned into an L1 and an L2 component. Unlike simple group mean aggregation, latent aggregation takes account of the heterogeneity within each group by partitioning the L1 variance σ^2_x from the L2 variance τ^2_x. In addition, by estimating latent group means at L2, which are assumed to cause the L1 observations in each group, the multilevel SEM approach acknowledges that the L1 scores do not perfectly map the construct at the L2 level, because of measurement error (Bennink et al., 2013, 2015; Preacher et al., 2010).

In sum, multilevel SEM replaces *manifest* with *latent* aggregation to aggregate individual-level predictors of macro-level outcomes. Like manifest aggregation, latent aggregation *per se* models only the transformational but not the situational and action formation mechanism. Statistically, however, latent aggregration is superior to manifest aggregation since it corrects for sampling error (see Table 1). As a result, its estimates are less biased, thereby permitting more valid inferences regarding the effect of multilevel predictors on contextual-level outcomes.

The two-step approach

The two-step approach also deals with the methodological and statistical issues that arise when studying multilevel models with contextual-level outcomes, albeit in a different manner. Figure 4 summarizes its basic idea.

The two-step approach builds on standard MLA. For an L1 outcome Y_{ij} and L1 units i nested in L2 contexts j, the standard model is given by:

$$Y_{ij} = \beta_{0j} + \beta_{1j} X_{ij} + e_{ij} \tag{4}$$

In equation (4), β_{0j} is the regression intercept of the outcome variable, β_{1j} is the regression slope of an L1 predictor, and e_{ij} is the residual error term. In contrast to non-nested regression analysis, both random intercepts β_{0j} and random slopes β_{1j} can be estimated for each L2 unit j by modeling them as a function of an additional L2 predictor Z_j with distinct intercepts (γ_{00} and γ_{10}) and regression slopes (γ_{01} and γ_{11}):

$$\beta_{0j} = \gamma_{00} + \gamma_{01} Z_j + u_{0j} \tag{5}$$

$$\beta_{1j} = \gamma_{10} + \gamma_{11} Z_j + u_{1j} \tag{6}$$

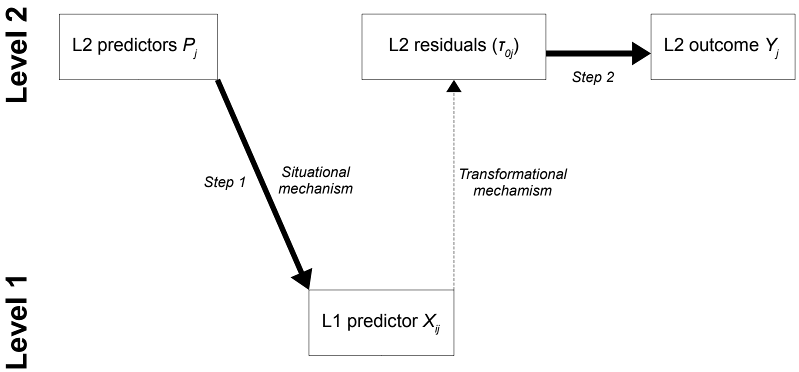


Figure 4 The two-step approach

Equations (5) and (6) introduce two additional residual error components: u_{0j} denotes the residual error of the outcome's L2 intercept β_{0j}, and u_{1j} denotes the residual error of the slope β_{1j} between L2 units.

Standard MLA only considers the case of an L1 outcome Y_{ij} that is predicted by L1 and L2 variables X_{ij} and Z_j, respectively. Griffin (1997) proposes an extension of the standard MLA approach to study an L2 outcome Y_j: Let X_{1ij} be the L1 explanatory variable of primary interest. In a first step, X_{1ij} is regressed on all other L1 and L2 predictors X_{2ij}, …, X_{nij} and Z_j:

$$X_{1ij} = \gamma_{00} + \gamma_{01} Z_j + \tau_{0j} + \beta_{1j} X_{2ij} + \ldots + \beta_{nj} X_{nij} + e_{ij} \tag{7}$$

In a second step, the L2 residuals u_{0j} of this model are used as a predictor variable in an L2 regression of the L2 outcome of interest:

$$Y_j = \beta_0 + \beta_1 u_{0j} + e_j \tag{8}$$

The effect of u_{0j} on the L2 outcome Y can be interpreted as the aggregated effect of the L1 variable X_1, net of both L1 and L2 covariates X_2, …, X_n and Z.

The two-step approach has both statistical and methodological advantages when studying multilevel models with contextual-level outcomes (see Table 1). Statistically, it provides a better estimate than the group mean aggregate: u_{0j} is a model-based estimate of the L2 variance that is already net of the L1 variance. In addition, u_{0j} can be adjusted for other covariates at L1 and L2. This may save degrees of freedom and circumvent collinearity issues when using u_{0j} as a predictor in a subsequent L2 regression. Compared to the group means approach and the multilevel SEM approach, the crucial methodological advantage of the two-step approach is its capacity to empirically model theoretical macro-micro-macro

Table 1 Comparison of methods for analyzing macro-micro-macro models

	Main methodological advantages & disadvantages	Main statistical advantages & disadvantages
group mean aggregation	Transformational mechanism (via manifest aggregation and macro regression)	Simple to perform, but only valid if variance of L1 variable = 0
ML SEM	Transformational mechanism (via latent aggregation and macro regression)	Takes sampling error into account: reduction of estimator bias
2-Step	1st step: situational & action-formation mechanism (via MLA) 2nd step: transformational mechanism (via residuals and macro regression)	Residual reflects the net effect of the individual-level independent variable

explanations in their entirety. The MLA of step 1 maps both the situational and action formation mechanism through the regression of an L1 outcome on L1 and L2 predictors. Storing the L2 residuals of this MLA then maps an underlying transformational mechanism in terms of an L1-L2 aggregation.

The relative statistical performance of each method can also be compared empirically. Based on previous research, we deduce two hypotheses. First, we expect that unless the L1 variance equals zero, simple group mean aggregation yields unreliable measures of the unobserved true group means. By contrast, multilevel SEM results in reliable estimates of true group means. Consequently, when group means based on simple aggregation are used as predictors of an L2 outcome, estimates of their regression coefficients may be biased (Bennink et al., 2013, 2015):

> H_1: Regression coefficients of L2 predictors that are simple group means deviate in terms of a) point estimates, b) standard errors, and c) resulting significance levels from regression coefficients of L2 predictors that have been aggregated through multilevel SEM.

Second, while the statistical performance of the two-step approach (Griffin, 1997) is less well researched, Lüdtke et al. (2008) compared multilevel SEM to another two-step approach proposed by Croon and van Veldhoven (2007). This approach adjusts the observed group means with weights from ANOVA formulas. This is quite similar to the decomposition of variance in an empty multilevel model. Lüdtke et al. (2008) observed that Croon and van Veldhoven's (2007) approach performed slightly less well than multilevel SEM. Consequently, we expect Griffin's two-step approach to yield estimates closer to multilevel SEM than to the simple group means approach:

> *H_2:* Regression coefficients of L2 predictors that have been aggregated by the two-step approach deviate less from multilevel SEM in terms of a) point estimates, b) standard errors, and c) resulting significance levels than regression coefficients of L2 predictors that are simple group means.

Substantive Application: A Multilevel Explanation of the Persistence of Democracy

Theoretical Background

To illustrate the methodological and statistical issues described in the previous section, we use the persistence of democracy as a substantive example. Explanations of democratic persistence pertain either to a macro-to-micro mechanism leading from the macro level to the level of individual citizens or to a micro-to-macro mechanism leading from individual citizens to the persistence of democracy at the macro level.

Przeworski (1991) introduces a classic model linking macro-level causes to individuals' micro-level incentives for subverting a democratic regime. Acknowledging that democratic competition produces winners and losers, he argues that "political forces comply with present defeats because they believe that the institutional framework that organizes the democratic competition will permit them to advance their interests in the future" (Przeworski, 1991, p. 19). Institutions are not only crucial for inspiring the belief that there will be future possibilities to advance one's interests. The given set of political and economic institutions also has distributional consequences affecting the capacities individuals have at their disposal to advance their interests (Przeworski, 1991). A model of democratic persistence therefore has to take into account that – under the same set of democratic rules – members of some societal groups might deem their chances of affecting future democratic outcomes to be lower than members of other societal groups. Correspondingly, classic studies have analyzed the decisive impact of economic development on both the process of successful democratization (Bollen, 1979; Bollen & Jackman, 1985; Lipset, 1959) as well as democratic persistence (Przeworski, Alvarez, Cheibub, & Limongi, 2000).

A second example for the macro-to-micro mechanism underlying the persistence of democracy is the idea that an ethnically divided society poses a particular challenge to democratic persistence (Horowitz, 1985; Rabushka & Shepsle, 1972; Reilly, 2001). In countries where several ethnic groups are politically mobilized, the question of who is to legitimately take part in the democratic game is continuously contested. Members of ethnic minorities often see little incentive to support ruling elites, who are – in virtue of the majority principle – likely to be members of the majority group. As a result, those out of power may choose to subvert democracy because they feel permanently excluded from democratic decisions likely to reflect only the interests of the majority.

A classic example of the micro-to-macro mechanism underlying the persistence of democracy is the political culture model. Almond and Verba (1963) seminally argued that the persistence of a political regime does not rest on its formal democratic institutions alone, but also on its political culture. Succeeding studies further specified the content of political culture and its effect on democratic persistence based on Easton's (1965, 1975) systems theory (Dalton, 2004; Fuchs, 2007; Norris, 1999). According to Easton, citizens' political support refers to their supportive values and attitudes toward the political community, the political regime, and political authorities (Easton, 1965). A critical amount of political support is necessary for any kind of political system to persist. Citizens' political support increases the functionality of political systems as it allows political authorities to convert demands into outputs and permits them to implement collectively binding decisions without having to resort to force (Easton, 1965).

Building on Easton (1965, 1975), Fuchs (2007) clarifies the implications of the different dimensions of political support for democratic political regimes. Support for the political authorities is crucial for their re- or de-election; support for the political system is essential for the persistence of a given type of democracy; support for democratic values is critical for the persistence of democracy in general (Fuchs, 2007). Thus, citizens' support for democratic values is the key factor when studying the effect of individual-level political orientations on the persistence of democracy at the macro level.

Fails and Pierce (2010) tested the systems approach of the political culture model empirically. Their analysis yielded no significant relationship between citizens' support for democratic values and their rejection of authoritarian values on the one hand and the probability of a decline of democracy on the other hand.

These mechanisms can be combined into a full multilevel explanation of democratic persistence (see Figure 5). From the macro to micro explanations, we take the insight that citizens' support for democratic values is likely to be affected by context-specific economic conditions and ethnic heterogeneity. From the micro to macro explanations, we take the insight that micro-level support for democratic values crucially accounts for the persistence of democracy at the macro level.

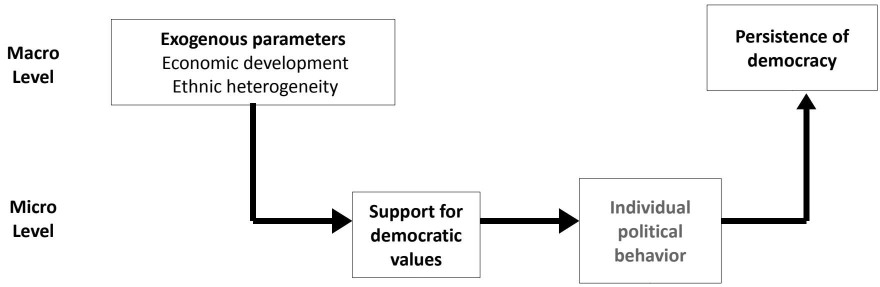


Figure 5 A two-level explanation of the persistence of democracy

Research Design

Period of analysis and data

Based on the data available, we analyzed the persistence of democracy from 1946 to 2014. We derived all L2 indicators from the *Quality of Government* standard time series data set (QoG) (Teorell et al., 2016), which includes data on a broad range of country-level indicators over time that we could easily merge with our L1 data.

To measure our L2 outcome variable – *democratic persistence* – we used the democracy index developed by the Polity IV project as included in the QoG (Marshall, Gurr, & Jaggers, 2015). Polity IV's democracy index – POLITY – reports countries' level of democracy on a scale ranging from -10 (fully autocratic) to +10 (fully democratic).[5] In line with the threshold provided on the Polity IV website (Marshall & Gurr, 2014), we considered countries as democracies if their POLITY score ≥ 6.[6]

As for our L2 predictors, we used the following indicators: *Economic development* was measured using countries' annual gross domestic product (GDP). We used the log of the OECD measure of GDP *per capita*. *Ethnic heterogeneity* was

5 POLITY is a composite score that quantifies the extent to which a country exhibits democratic and authoritarian characteristics. Polity IV coders assess countries' formal political institutions in terms of five component variables – the competitiveness of political participation (1), the openness of executive recruitment (2), the competitiveness of executive recruitment (3), the constraints on the executive (4), and the regulation of political participation (5) for each country on an annual basis. Countries are assigned weighted scores for each component. These are then added up to arrive at a democracy (DEMOC) and an autocracy score (AUTOC), both of which range from 0 to 10. The autocracy score is then subtracted from the democracy score to construct POLITY (Marshall et al., 2015).

6 We noted an inconsistency in the definition of the thresholds. In their codebook, Marshall et al. (2015) state that POLITY values ranging from +7 to +10 indicate a democratic regime.

measured using Fearon's (2003) ethno-linguistic fractionalization index (ELF), a measure of the probability that two randomly chosen individuals from a particular country are members of different ethnic groups. It ranges from 0 (perfect homogeneity) to 1 (very high fractionalization).[7]

Citizens' support for democratic values and all other L1 covariates were derived from the *World Values Survey* (WVS). The WVS is a cross-national survey based on representative national samples investigating worldwide socio-cultural and political change. For our analyses, we used the wave 6 aggregated longitudinal file, which includes more than 340,000 observations sampled in 101 countries across all available waves from 1981 to 2014. In line with previous research, *support for democratic values* was operationalized in terms of respondents' reply to the following question: "I'm going to describe various types of political systems and ask what you think about each as a way of governing this country. For each one, would you say it is a very good, fairly good, fairly bad or very bad way of governing this country?". For reasons of data availability, we used respondents' rejection of an authoritarian system rather than their support for a democratic system. The answer category reads: "Having a strong leader who does not have to bother with parliament and elections" (1 = 'very good'; 2 = 'fairly good'; 3 = 'bad'; 4 = 'very bad'). For our analyses, we dichotomized this variable (0 = 'good / very good' vs. 1 = 'bad / very bad'). In accordance with previous research (Schneider, 2009), we controlled for individuals' age (six categories ranging from 1 = '15-24 years' to 6 = '65 and more years'), subjective assessment of social class (five categories ranging from 1 = 'lower class' to 5 = 'upper class'), and education (eight categories ranging from 1 'inadequately completed elementary education' to 8 'university with degree/higher education').[8]

Methods of analysis

Studying the effect of L1 and L2 predictors on an L2 outcome such as the persistence of democracy poses two methodological challenges. First, choosing a method to address the L1-L2 aggregation problem; second, analyzing persistence of democracy, which is a duration variable.

We compared three different strategies for solving the L1-L2 aggregation problem. First, we aggregated support for democratic values and all other L1 covariates by computing the arithmetic means for each country year (model 1). Second, we corrected for sampling error by estimating a latent aggregation of all L1 variables on L2 using multilevel SEM (model 2).[9] Third, we applied the two-step procedure

7 The formula is: $1 - \sum_{i=1}^{n} s_i^2$ where s_i is the share of group i ($i = 1, \ldots, n$).

8 See Table A1 (appendix) for a summary of all variables.

9 The latent aggregation was performed in Mplus, Version 7 (Muthén & Muthén, 2012).

proposed by Griffin (1997) by regressing support for democratic values on all other L1 and L2 predictors and then using the L2 residuals of this multilevel model as a new predictor variable.

We estimated not one, but several multilevel levels that were built up stepwise: The first empty model separated the L2 residuals of support for democratic values from the L1 residuals (model A1). We then added the macro level predictors GDP and ELF (models A2-A4). Finally, we added all L1 controls (model A5).[10] Researchers typically use stepwise model building (which we also carried out in the L2-only regressions below) to make causal claims about mediator variables partialing out significant effects of previous regressors. Apart from comparing point estimates and confidence intervals between aggregation methods for the final model, we also considered it instructive to analyze a series of stepwise models in order to assess whether different aggregation methods lead to different claims about causal mediation.

In addition, we chose an adequate model for predicting democratic persistence, a duration variable. The time span of interest is the persistence of a given democracy until its breakdown. Whereas some democracies may have persisted before entering the observation window (left censoring), others may have continued to persist after the observation ended (right censoring). Within the time period of analysis, the same country may have experienced multiple democratic sequences, followed by breakdowns. In order to address these issues, we used event history modeling. We considered democratic breakdown to occur if the score of democratic regimes (nested within countries) fell below the threshold of POLITY = 6. The duration until this event was measured by the total number of years a democratic system persisted from 1946 onwards. Multiple breakdowns within the same country were coded as distinct events. To keep the models parsimonious, we used a simple exponential event history model, which assumes constant transition rates across years.

In formal terms, our event history model is defined as follows: Let h denote the hazard rate of democracies' estimated risk of falling below POLITY = 6 and t the time of democracies' survival. The basic exponential survival model can then be described as:

$$h(t) = \lambda; t > 0, \lambda > 0 \tag{9}$$

λ is a positive constant constraining transition rate (in terms of democratic breakdowns) that is equal across years. Our aim was to predict the expected survival time E(t) with an aggregate measure of citizens' support for democratic values (*DVAL*),

10 See Table A2 (appendix).

countries' *GDP* and *ELF,* as well as aggregate measures of citizens' age (*AGE*), subjective social class (*SCLASS*), and education (*EDUC*).

When applying simple aggregation, democracies' expected time of survival was estimated by:

$$E\left(t_j\right) = \exp\begin{pmatrix} \beta_0 + \beta_1 \overline{DVAL_{\bullet j}} + \beta_2 GDP_j + \beta_3 ELF_j + \beta_4 \overline{AGE_{\bullet j}} \\ + \beta_5 \overline{SCLASS_{\bullet j}} + \beta_6 \overline{EDUC_{\bullet j}} \end{pmatrix} \tag{10}$$

where $\overline{X_{\bullet j}}$ from equation (1) was replaced by the aforementioned predictor variables. When using latent aggregation, we estimated:

$$E\left(t_j\right) = \exp\begin{pmatrix} \beta_0 + \beta_1 U\left(DVAL\right)_j + \beta_2 GDP_j + \beta_3 ELF_j + \beta_4 U\left(AGE\right)_j \\ + \beta_5 U\left(SCLASS\right)_j + \beta_6 U\left(EDUC\right)_j \end{pmatrix} \tag{11}$$

Here, *U* refers to the unobserved latent L2 group mean which is assumed to cause the observed L1 values of each variable.

Finally, when employing the two-step approach, the estimates were derived as follows:

$$E\left(t_j\right) = \exp\left(\beta_0 + \beta_1 u_{0jm}\right) \tag{12}$$

In equation (12), u_{0jm} denotes the L2 residuals from a hierarchical regression of citizens' support for democratic values on both the L2 predictors and the L1 covariates. The subscript *m* indicates that the hierarchical models were built up in a stepwise manner, which is why we estimated several terms for u_0.

These formal specifications require a methodological addendum: While we *estimated* three *L2* event history analyses after having applied each of the three aggregation methods, our *theoretical* explanation emphasizes the importance of citizens' support for democratic values on *L1*. Hence, though the event history models applied L2-only regressions, in line with the paradigm of structural individualism, we assume that the theoretical mechanisms operate via citizens' preferences and beliefs on the micro level. In line with the aim of our article, we sought to determine how the three different aggregation methods map these L1 processes when predicting an L2 outcome.

In order to increase our statistical power, we used both inter- and extrapolation techniques for our independent variables. We interpolated missing values between observation points, using the *-ipolate-* command in Stata. In addition, we extrapolated missing values between the last valid observation and 2015, using a 'non-linear trend' scenario. We first estimated a polynomial regression of the interpolated values of each predictor on years of observations using the *-lpoly-* command in

Stata. We then used out-of-sample predicted values to replace missing observation for subsequent years over countries.[11]

Results

Prior to computing the comprehensive multivariate models, we compared the survival functions of democracies with high vs. low average support for democratic values. We dichotomized the support variable and compared countries with one standard deviation above vs. below the grand mean of the aggregated variable. We then compared the survival functions of these two groups of countries using group mean aggregation, the two-step model, and latent aggregation. Independent of the method of aggregation, in the long run, the estimated survival rate for democracies scoring one standard deviation above the grand mean of support for democratic values was higher than for their lower-scoring counterparts (see Figure A3, appendix). Apart from a lower estimate of the survival rate of countries whose citizens had less support for democratic values in the two-step model, the differences between the aggregation methods appeared to be negligible.

Figure 6 presents the results of the analyses using the simple group means approach (model 1), multilevel SEM (model 2), and the two-step approach (model 3). It shows both point estimates and confidence intervals for the L1 and L2 predictors. Our survival models were built up stepwise: In model 1a and 2a, the survival rate of democracies was first predicted by support for democratic values only; in model 3a, it was predicted by the L2 residuals from the multilevel null model, which separated the variance of the L1 support variable without having included any other L1 or L2 predictor. In models 1b and models 2b, we simultaneously added GDP and ELF. Correspondingly, in model 3b we included the residuals corrected for these L2 predictors. Finally, in model 1c and 2c, we added the L1 covariates; in model 3c we included the residuals corrected for the L1 covariates. Because of the low number of events, we displayed confidence intervals both on the 10% ($|t| > 1.64$; see ticks of confidence bands) and the 5% significance level ($|t| > 1.96$; see ends of confidence bands).

When applying the *simple group means approach*, support for democratic values did not turn out to be a significant predictor of democratic survival. Point estimates varied between -3.734 in model 1a and -3.367 in model 1c, but neither

11 The overlap of valid observations for both democratic persistence and support for democratic values before and after interpolation is displayed in Figure A1 (appendix). The basic survivor function of democratic persistence for our reduced sample of analysis is sufficiently similar to the survivor function of the total country sample (see Figure A2, appendix). As a sensitivity check, we also extrapolated our interpolated values by repeating the last valid observation of each predictor for subsequent years with missing values. Results based on this extrapolation technique are very similar to the results reported in the results section (see Figure A4, appendix).

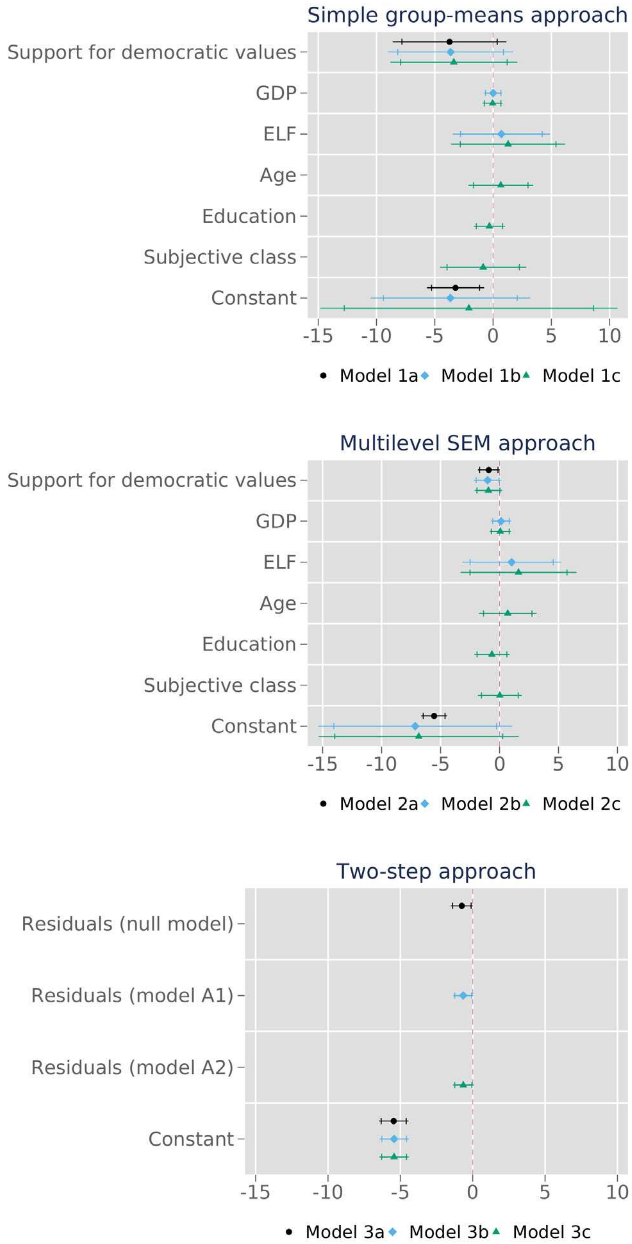


Figure 6 Point estimates and confidence intervals of countries' democratic survival across aggregation methods. *N*= 917 observations, *N*= 122 subjects, *N*= 5 failures in all models

estimate was larger than 1.65 times its standard error (also see Table A3, appendix). The latter also applies to all other L2 predictors and to the L1 covariates. We observed significant intercept variation in model 1a, which only included support for democratic values as a predictor variable, but not in models 1b and 1c, which controlled for the remaining variables. Values of AIC and BIC as indicators of model fit show that not much was gained by adding predictors of democratic survival apart from citizens' support for democratic values (see Table A3, appendix).

When using the *latent aggregation approach*, the estimated confidence intervals of support for democratic values became more precise and we observed two effects of support for democratic values on democratic survival that were greater than 1.65 times their standard error (models 2a and 2b). Once the aggregated L1 covariates were controlled for, our predictor was no longer significantly associated with the outcome. Point estimates were remarkably lower after latent aggregation, ranging from -.911 in model 2a to -1.009 in model 2b (see Table A4, appendix). Having controlled for L2 structural conditions (in terms of GDP and ELF), the effect of support for democratic values became more negative from model 2a to model 2b – which points to a suppressor effect. Yet, similar to the simple group means analysis, none of the remaining variables turned out to be significant predictors of democratic survival. Model fit indices again supported the most parsimonious model 2a and intercept variation was significant in the first two submodels only.

When applying the *two-step approach*, point estimates of support for democratic values on democratic stability were predicted with similar precision as in latent aggregation when looking at the confidence intervals. Yet, in the two-step model, we observed three significant effects at the 10% level. The L2 (u_{oj}) residuals of support for democratic values predicted democratic survival independent of whether they were adjusted for other L1 or L2 variables. Effect sizes ranged from -.754 in model 3a to -.651 in model 3c (see Table A5, appendix). In contrast to simple group mean and latent aggregation, the intercept remained significant in all three sub-models. Though model fit indices supported the most parsimonious model 3a, the differences between model fit indices across models were less striking than in the event history regressions following manifest and latent aggregation.

Our results can be summarized as follows: In each estimation, support for democratic values was negatively associated with the event of democratic breakdown, as expected by theory. This replicated our bivariate analysis where democracies with higher support for democratic values showed a longer estimated survival rate on average. Apart from this similarity, there are notable differences between the aggregation methods: While support for democratic values was not significantly associated with democratic stability after manifest aggregation, significant effects could be observed after both latent aggregation and the two-step approach. Applying more advanced aggregation methods led to smaller point estimates and standard errors compared to the simple group means approach. All this is in line with the two hypotheses postulating notable differences between simple group means aggregation and latent aggregation, and closer similarity between the two-step approach and latent aggregation than between the two-step approach and manifest aggregation.

Yet, compared to latent aggregation, which has already been observed to yield unbiased point estimates in simulation models (Bennink et al., 2013, 2015; Lüdtke et al., 2008), researchers who apply the two-step approach may run the risk of committing type one errors: In the most comprehensive model of the two-step approach (model 3c) and unlike in the corresponding regressions following latent aggregation (model 2c), the effect of support for democratic values was significant at the 10% level.[12]

Conclusion

In this paper, we addressed a methodological challenge well known to comparative survey researchers: how to study the effect of level two (L2) and level one (L1) predictors of a level two (L2) outcome so as to yield both reliable and valid results. Researchers have criticized simple aggregation for methodological and statistical reasons. Building on these insights and using the persistence of democracy as a substantive example, we compared the simple group means approach with two more advanced analytical strategies: the multilevel SEM approach, which estimates a latent L2 variable assumed to cause its L1 indicators, and a two-step approach, which relies on the L2 residuals of a multilevel model estimated prior to the analysis of interest (Griffin, 1997).

Our study corroborates previous critiques of the simple group-means approach. In both bivariate comparisons of countries' survival curves and more comprehensive multivariate event history analyses, we observed that support for democratic values was negatively associated with democratic breakdown. Unlike in the bivariate models, however, the multivariate models revealed that the associated significance levels of the estimates of support for democratic values differed remarkably depending on the aggregation method. Whereas support for democratic values was not significant in the regressions following simple group mean aggregation, confidence intervals suggested point estimates of higher precision when using either the multilevel SEM or the two-step approach, and the latter two approaches showed several significant effects at the 10% level.

These empirical results show that researchers can improve the validity of their inferences by choosing more advanced analytical strategies. First, the results match previous findings from simulation analyses (Lüdtke et al., 2008), which show that

12 The event-history models underlying Figure 6 are listed in Tables A3 to A5 (appendix).

the simplest form of aggregation – manifest group means – is prone to beta or type-two errors in terms of false negative findings. Second, our results challenge Fails and Pierce's (2010) finding (based on simple aggregation) that support for democratic values has no effect on democracies' probability of decline. Our results suggest that comparative survey researchers interested in the effect of one or more L1 predictors on an L2 outcome may overestimate the standard errors of their regression coefficients when using manifest group mean aggregation.

The two more advanced analytical strategies have distinct methodological and statistical advantages. From a statistical perspective, the two-step approach performs somewhat poorer than the multilevel SEM approach: Given that simulation revealed regression coefficients after latent aggregation to be unbiased (Bennink et al., 2013, 2015; Lüdtke et al., 2008), researchers who apply the two-step approach may run the risk of committing type-one errors in terms of false positive findings. An evident methodological advantage of the two-step approach is, however, that it is particularly suited to simultaneously model situational, action formation, and transformational mechanisms in their entirety.

We conclude with several suggestions for future research. As of yet, no simulation analyses (similar to the ones comparing the simple group mean and the multilevel SEM approach) have been carried out for the two-step approach. It is therefore not possible to determine whether the estimated confidence intervals of the two-step approach are more or less reliable than the results of the latent aggregation approach. Hence, our first suggestion for future research is to perform a simulation analyses for all three aggregation methods. Controlling the data-generating mechanism would permit valid conclusions about the actual precision of each aggregation method compared to the 'real' effect size at L2.

Second, the latent aggregation model can be extended towards a *doubly-latent* model with controls for measurement error. Thus, our second suggestion for future research is to use multiple indicators of political support to arrive at a doubly-latent model of political support at L2. Depending on the results of the aforementioned simulation study, latent variable models and the two-step approach could eventually also be combined in order to estimate both situational and transformational mechanisms without falling prey to either measurement or sampling error. Moreover, if individuals' actual decisions such as turning out to vote or participating in demonstrations or public protests are considered, a combined framework of structural equation modeling and the two-step approach would allow researchers to map action-formation mechanisms as well.[13] Third, while we used a simple exponential event-history model to simplify the analysis, future research might make use of

13 Structural equation modeling can map action formation mechanisms in simple L1 regressions as well. In addition, for group-mean centered L1 variables, multilevel SEM can estimate situational mechanisms by computing the difference between L2 and L1 regression coefficients (Marsh et al., 2009).

more flexible links for the survival function such as piecewise constant or frailty models.

In sum, we encourage comparative survey researchers to surpass the simple group means aggregation approach in favor of more advanced methods of analyzing contextual-level outcomes. We have shown that this helps researchers to circumvent beta or type-two errors in terms of false negative findings when using one or more L1 indicator to predict an L2 outcome. In addition, unlike the simple group means approach, these more advanced methods can be extended further, thereby facilitating the test of more theoretically valid models.

References

Almond, G. A., & Verba, S. (1963). *The civic culture: Political attitudes and democracy in five nations*. Princeton, NJ: Princeton University Press.

Becker, D., Beckers, T., Franzmann, S. T., & Hagenah, J. (2016). Contextualizing cognitive consonance by a social mechanisms explanation: Moderators of selective exposure in media usage. *Analyse & Kritik*, *38*(1), 149-178. http://doi.org/10.1515/auk-2016-0108

Bennink, M., Croon, M. A., & Vermunt, J. K. (2013). Micro-macro multilevel analysis for discrete data: A latent variable approach and an application on personal network data. *Sociological Methods & Research*, *42*(4), 431-457. https://doi.org/10.1177/0049124113500479

Bennink, M., Croon, M. A., & Vermunt, J. K. (2015). Stepwise latent class models for explaining group-level outcomes using discrete individual-level predictors. *Multivariate Behavioral Research*, *50*(6), 662-675. http://dx.doi.org/10.1080/00273171.2015.1074879

Bollen, K. A. (1979). Political democracy and the timing of development. *American Sociological Review*, *44*(4), 572-587. Retrieved from http://www.jstor.org/stable/2094588

Bollen, K. A., & Jackman, R. W. (1985). Economic and noneconomic determinants of political democracy in the 1960s. In R. G. Braungart & M. M. Braungart (Eds.), *Research in political sociology* (pp. 27-48). Greenwich, CT: JAI Press.

Bryk, A. S., & Raudenbush, S. W. (1992). *Hierarchical linear models: Applications and data analysis methods*. Newbury Park, CA: Sage.

Croon, M. A., & van Veldhoven, M. J. P. M. (2007). Predicting group-level outcome variables from variables measured at the individual level: A latent variable multilevel model. *Psychological Methods*, *12*(1), 45-57. http://doi.org/10.1037/1082-989X.12.1.45

Dalton, R. J. (2004). *Democratic challenges, democratic choices: The erosion of political support in advanced industrial democracies*. Oxford, UK: Oxford University Press.

Diamantopoulos, A., & Winklhofer, H. M. (2001). Index construction with formative indicators: An alternative to scale development. *Journal of Marketing Research*, *38*(2), 269-277.

Easton, D. (1965). *A systems analysis of political life*. New York, NY: Wiley.

Easton, D. (1975). A re-assessment of the concept of political support. *British Journal of Political Science*, *5*(4), 435-457.

Fails, M. D., & Pierce, H. N. (2010). Changing mass attitudes and democratic deepening. *Political Research Quarterly*, *63*(1), 174-187. http://doi.org/10.1177/1065912908327603

Fearon, J. D. (2003). Ethnic and cultural diversity by country. *Journal of Economic Growth*, *8*(2), 195-222. http://doi.org/10.1023/A:1024419522867

Fuchs, D. (2007). Political culture paradigm. In R. J. Dalton & H.-D. Klingemann (Eds.), *The Oxford handbook of political behavior* (pp. 161–184). Oxford, UK: Oxford University Press.

Griffin, M. A. (1997). Interaction between individuals and situations: Using HLM procedures to estimate reciprocal relationships. *Journal of Management*, *23*(6), 759-773. http://doi.org/10.1016/S0149-2063(97)90028-3

Hedström, P., & Swedberg, R. (Eds.). (1998). *Social mechanisms: An analytical approach to social theory*. Cambridge, UK: Cambridge University Press.

Hedström, P., & Ylikoski, P. (2010). Causal mechanisms in the social sciences. *Annual Review of Sociology*, *36*, 49-67. https://doi.org/10.1146/annurev.soc.012809.102632.

Horowitz, D. L. (1985). *Ethnic groups in conflict*. Berkeley, CA: University of California Press.

Hox, J. J. (2010). *Multilevel analysis: Techniques and applications* (2nd ed.). New York, NY: Routledge.

Lim, F., Bond, M. H., & Bond, M. K. (2005). Linking societal and psychological factors to homicide rates across nations. *Journal of Cross-Cultural Psychology*, *36*(5), 515-536. http://doi.org/10.1177/0022022105278540

Lipset, S. M. (1959). Some social requisites of democracy: Economic development and political legitimacy. *The American Political Science Review*, *53*(1), 69-105. http://doi.org/10.2307/1951731

Lüdtke, O., Marsh, H. W., Robitzsch, A., Trautwein, U., Asparouhov, T., & Muthén, B. O. (2008). The multilevel latent covariate model: A new, more reliable approach to group-level effects in contextual studies. *Psychological Methods*, *13*(3), 203-229. http://doi.org/10.1037/a0012869

Marsh, H. W., Lüdtke, O., Robitzsch, A., Trautwein, U., Asparouhov, T., Muthén, B. O., & Nagengast, B. (2009). Doubly-latent models of school contextual effects: Integrating multilevel and structural equation approaches to control measurement and sampling error. *Multivariate Behavioral Research*, *44*(6), 764-802. http://doi.org/10.1080/00273170903333665

Marshall, M. G., & Gurr, T. R. (2014). *Polity IV Project*. Retrieved from http://www.systemicpeace.org/polity/polity4.htm

Marshall, M. G., Gurr, T. R., & Jaggers, K. (2015). *Polity IV Project: Regime characteristics and transitions, 1800-2015. Dataset users' manual*. Vienna, Austria: Center for Systemic Peace. Retrieved from http://www.systemicpeace.org/inscr/p4manualv2015.pdf

Muller, E. N., & Seligson, M. A. (1994). Civic culture and democracy: The question of causal relationships. *American Political Science Review*, *88*(3), 635-652. http://doi.org/10.2307/2944800

Muthén, L. K., & Muthén, B. O. (2012). *Mplus User's Guide. Seventh Edition*. Los Angeles, CA: Muthén & Muthén.

Norris, P. (Ed.). (1999). *Critical citizens: Global support for democratic government*. Oxford, UK: Oxford University Press.

OECD. (2014). *Social expenditure: Reference series (Edition 2016)*. [Data file]. Retrieved from http://doi.org/10.1787/socx-data-en

Opp, K.-D. (2011). Modeling micro-macro relationships: Problems and solutions. *Journal of Mathematical Sociology*, *35*(1-3), 209-234. http://doi.org/10.1080/0022250X.2010.532257

Preacher, K. J., Zyphur M. J., & Zhang, Z. 2010. A general multilevel SEM framework for assessing multilevel mediation. *Psychological Methods 15*(3), 209-233. http://psycnet.apa.org/doi/10.1037/a0020141

Przeworski, A. (1991). *Democracy and the market: Political and economic reforms in Eastern Europe and Latin America*. Cambridge, UK: Cambridge University Press.

Przeworski, A., Alvarez, M. E., Cheibub, J. A., & Limongi, F. (2000). *Democracy and development: Political institutions and well-being in the world, 1950-1990*. Cambridge, UK: Cambridge University Press.

Rabushka, A., & Shepsle, K. A. (1972). *Politics in plural societies: A theory of democratic instability*. Columbus, OH: Merrill.

Raub, W., Buskens, V., & van Assen, M. A. L. M. (2011). Micro-macro links and microfoundations in sociology. *The Journal of Mathematical Sociology*, *35*(1-3), 1-25. http://doi.org/10.1080/0022250X.2010.532263

Reilly, B. (2001). *Democracy in divided societies: Electoral engineering for conflict management*. Cambridge, UK: Cambridge University Press.

Schneider, C. Q. (2009). *The consolidation of democracy: Comparing Europe and Latin America*. New York, NY: Routledge.

Shin, Y. & Raudenbush, S. W. 2010. A Latent Cluster-Mean Approach to the Contextual Effects Model With Missing Data. *Journal of Educational and Behavioral Statistics, 35*(1), 26-53. https://doi.org/10.3102/1076998609345252

Snijders, T. A. B., & Bosker, R. J. (1999). *Multilevel analysis: An introduction to basic and advanced multilevel modeling*. London, UK: Sage.

Teorell, J., Dahlberg, S., Holmberg, S., Rothstein, B., Khomenko, A., & Svensson, R. (2016). *The quality of government standard dataset (Version Jan16)*. [Data file]. University of Gothenburg: The Quality of Government Institute. http://doi.org10.18157/QoGStdJan16

Tranow, U., Beckers, T., & Becker, D. (2016). Explaining and understanding by answering “why” and “how” questions: A programmatic introduction to the special issue “social mechanisms.” *Analyse & Kritik*, *38*(1), 1-30. http://doi.org/10.1515/auk-2016-0102

Udehn, L. (2002). The changing face of methodological individualism. *Annual Review of Sociology*, *28*, 479-507. Retrieved from http://www.jstor.org/stable/3069250

World Values Survey Association. (2015). *World Values Survey 1981-2014 Longitudinal Aggregate (v. 20150418)*. [Data file]. Retrieved from http://www.worldvaluessurvey.org/WVSDocumentationWVL.jsp

Appendix

Table A1 Distribution of all indicators

		count	mean	sd	min	max
LEVEL 1	Support for democratic values	269869	2.75	1.03	1	4
	Support for democratic values (dichotomized)	269869	0.59	0.49	0	1
	Age recoded	337018	3.1	1.57	1	6
	Highest educational level attained	296142	4.72	2.23	1	8
	Subjective social class	284337	2.68	0.99	1	5
LEVEL 2	GDP	7998	7.62	1.64	3.51	12.11
	ELF	8573	0.47	0.27	0.00	1.00
	Support for democratic values	1007	0.60	0.18	0.01	0.97
	Age	1190	3.19	0.46	1.91	4.30
	Education	1076	4.74	0.80	2.53	6.79
	Subjective class	1022	2.69	0.28	1.70	3.69
	Residuals (null model)	921	-0.02	0.86	-4.84	3.06
	Residuals (model A1)	921	0.00	0.93	-5.34	2.89
	Residuals (model A2)	921	-0.01	0.95	-5.30	2.94
	Support for democratic values	1007	-0.02	0.82	-3.89	2.22
	Age	1058	0.07	0.53	-1.47	1.32
	Education	1034	0.09	0.65	-1.67	1.71
	Subjective class	1013	0.04	0.51	-1.67	1.35

Table A2 Multilevel logistic regression of support for democratic values (dichotomized) on level-two predictors and level-one covariates

	Null model		Model 1a		Model 1b	
	b	*se*	*b*	*se*	*b*	*se*
Intercept	1.812**	(0.585)	2.042***	(0.580)	0.457***	(0.071)
log(GDP)			-0.174**	(0.061)	-0.166**	(0.061)
ELF			-0.392	(0.345)	-0.427	(0.341)
Age: 15-24 years	REFERENCE CATEGORY					
25-34 years					0.015	(0.015)
35-44 years					0.067***	(0.015)
45-54 years					0.103***	(0.017)
55-64					0.092***	(0.019)
65 and more years					-0.039	(0.020)
Education: Inadequately completed elementary	REFERENCE CATEGORY					
Completed elementary					0.042	(0.022)
Incomplete secondary: tech./voc.					0.051*	(0.025)
Completed secondary: tech./voc.					0.178***	(0.022)
Incomplete secondary: univ. prep.					0.171***	(0.024)
Complete secondary: univ. prep.					0.274***	(0.022)
Some university without degree					0.428***	(0.026)
University with degree					0.581***	(0.023)
Subjective class: lower	REFERENCE CATEGORY					
working					0.016	(0.017)
lower middle					0.042*	(0.017)
upper middle					-0.034	(0.019)
upper					-0.275***	(0.038)
τ_{0j}	0.025	(0.063)	0.012	(0.063)	-0.045	(0.054)
N	219740		219740		219740	
AIC	261954		263445		263440	

Notes. Random intercept model (QR decomposition) across country-years (level 2). Significance levels: * < .05; ** < .01; *** < .001 (two-sided). Standard errors in parentheses.

Table A3 Exponential event-history regression of democratic breakdown on aggregated support for democratic values, L2 predictors, and aggregated L1 controls (simple group-means approach)

	Model 2a *b/se*	Model 2b *b/se*	Model 2c *b/se*
Intercept	-3.220*	-3.662	-2.073
	(1.252)	(3.492)	(6.503)
Support for democratic values	-3.734	-3.642	-3.367
	(2.485)	(2.754)	(2.783)
log(GDP)		0.01	-0.038
		(0.399)	(0.432)
ELF		0.715	1.294
		(2.131)	(2.495)
Age			0.662
			(1.419)
Education			-0.315
			(0.685)
Subjective class			-0.846
			(1.887)
AIC	43.318	47.201	52.375
BIC	52.96	66.486	86.123
N (failures)	5	5	5
N (subjects)	122	122	122
N (observations)	917	917	917

Notes. Significance levels: + < .10: * < .05; ** < .01; *** < .001 (two-sided). Standard errors in parentheses.

Table A4 Exponential event-history regression of democratic breakdown on aggregated support for democratic values, L2 predictors, and aggregated L1 controls (multilevel SEM approach)

	Model 3a	Model 3b	Model 3c
	b/se	*b/se*	*b/se*
Intercept	-5.547***	-7.151+	-6.851
	(0.563)	(4.195)	(4.332)
Support for democratic values	-0.911+	-1.009+	-0.945
	(0.474)	(0.592)	(0.591)
GDP		0.132	0.064
		(0.428)	(0.461)
ELF		1.029	1.611
		(2.141)	(2.502)
Age			0.696
			(1.249)
Education			-0.644
			(0.769)
Subjective class			0.024
			(0.949)
AIC	42.444	46.179	51.203
BIC	52.086	65.463	84.951
N (failures)	5	5	5
N (subjects)	122	122	122
N (observations)	917	917	917

Notes. Significance levels: + < .10: * < .05; ** < .01; *** < .001 (two-sided). Standard errors in parentheses.

Table A5 Exponential event-history regression of democratic breakdown on residualised support for democratic values (two-step approach)

	Model 4a *b/se*	Model 4b *b/se*	Model 4c *b/se*
Intercept	-5.460*** (0.525)	-5.427*** (0.517)	-5.433*** (0.520)
Residuals (Null model)	-0.754+ (0.389)		
Residuals (model 1a)		-0.658+ (0.357)	
Residuals (model 1b)			-0.651+ (0.361)
AIC	42.813	43.047	43.089
BIC	52.455	52.689	52.731
N (failures)	5	5	5
N (subjects)	122	122	122
N (observations)	917	917	917

Notes. Significance levels: + < .10: * < .05; ** < .01; *** < .001 (two-sided). Standard errors in parentheses.

Figure A1 Distribution of democratic persistence and support for democratic values across country years

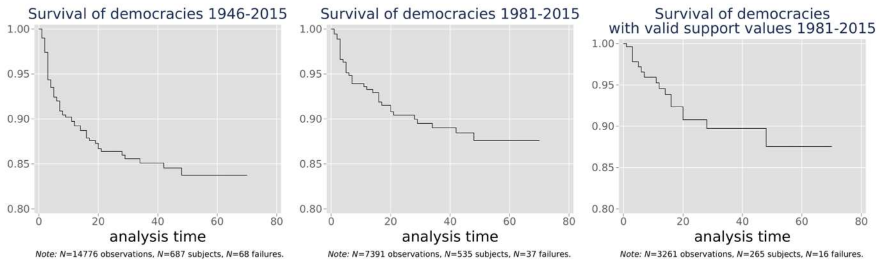


Figure A2 A comparison of democracies' estimated survival rates across different samples of analysis

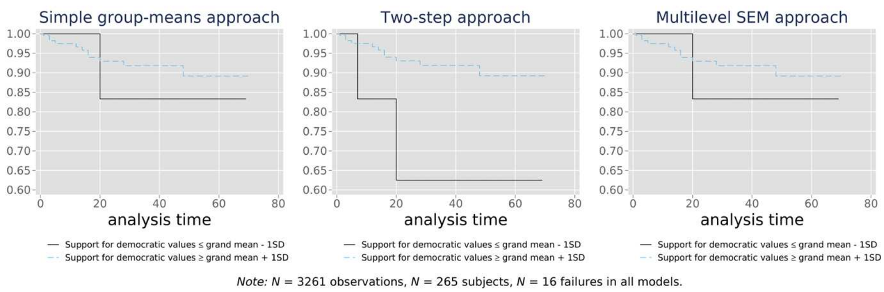


Figure A3 Survival of democracies by support for democratic values across aggregation methods

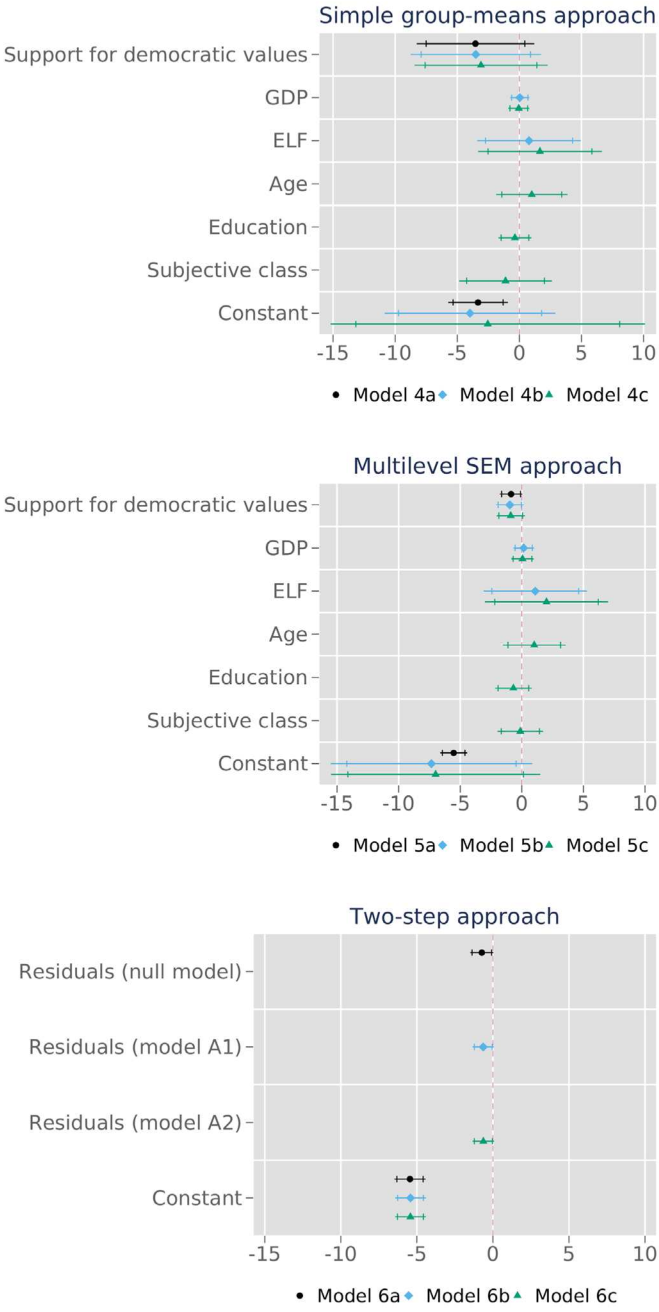


Figure A4 Point estimates and confidence intervals of countries' democratic survival across aggregation methods (constant interpolation)

methods, data, analyses | Vol. 12(2), 2018, pp. 265-308 DOI: 10.12758/mda.2017.07

Simultaneous Feedback Models with Macro-Comparative Cross-Sectional Data

Nate Breznau
Mannheim Centre for European Social Research (MZES)

Abstract

Social scientists often work with theories of reciprocal causality. Sometimes theories suggest that reciprocal causes work simultaneously, or work on a time-scale small enough to make them appear simultaneous. Researchers may employ simultaneous feedback models to investigate such theories, although the practice is rare in cross-sectional survey research. This paper discusses the certain conditions that make these models possible if not desirable using such data. This methodological excursus covers the construction of simultaneous feedback models using a structural equation modeling perspective. This allows the researcher to test if a simultaneous feedback theory fits survey data, test competing hypotheses and engage in macro-comparisons. This paper presents methods in a manner and language amenable to the practicing social scientist who is not a statistician or matrix mathematician. It demonstrates how to run models using three popular software programs (*MPlus*, *Stata* and *R*), and an empirical example using *International Social Survey Program* data.

Keywords: simultaneous feedback model, cross-sectional data, macro-comparative research, structural equation modeling, reciprocal causality, *Mplus*, *Stata*, *R* (lavaan)

Social scientists often study reciprocally causal phenomena. For example, supply and demand in economics; candidate evaluations and party identification in political science; road investment and travel demand in geography; and educational attainment and parenthood entry in sociology and demography (Marini, 1984; Page & Jones, 1979; Xie & Levinson, 2010). When timings of reciprocal causes are unobservable or occur contemporaneously, a state of simultaneous feedback exists. Rather than in cycles, events happen at the same time. Philosophers of causality question the existence of simultaneous feedback (Mulaik, 2009: Chapter 3); however, researchers regularly face theoretical and data conditions that force them to accept simultaneous feedback in practice. This is particularly acute in macro-comparative survey research where observations take place over a year, but theoretical causes may take place at less-than-yearly intervals. All sub-yearly causal effects appear simultaneous within a year interval. Under certain conditions, macro-comparative researchers can employ simultaneous feedback models (SFMs) to capture these effects, allowing them to overcome some limitations of comparative cross-sectional survey research.

Herein, I elaborate when and how to use SFMs. This requires structural equation modeling (SEM) strategies to explicate theoretical relationships before extracting meaningful statistical results. I use minimal statistical and mathematical jargon without matrix algebra[1], and a practical example of public opinion and social policy. I show that SFMs provide a powerful method for macro-comparative survey researchers to explain, predict and compare reciprocally causal phenomena.

Simultaneous Feedback

Instances where two phenomena are co-causes of each other are ubiquitous in social research[2]; however, modeling reciprocal causality is challenging. Time is usually

1 Matrix algebra is the basis of nearly all social science statistics including SFMs; however, this excursus is for the practicing social scientist who is unlikely a matrix algebraician.

2 More non-exhaustive examples: (Brehm & Rahn, 1997; Chong & Gradstein, 2007; Claibourn & Martin, 2000; Liska & Reed, 1985; Mulatu & Schooler, 2002; Owens, 1994; Thornton, Axinn, & Hill, 1992)

Acknowledgments

I am grateful to two anonymous reviewers, discussions on SEMNET, Bart Meuleman and Sebastian Pink for helpful instruction and comments.

Direct correspondence to

Nate Breznau, Mannheim Centre for European Social Research (MZES)

E-Mail: breznau.nate@gmail.com

the basis for explaining or predicting things (Elwert, 2013; Pedhazur, 1997). To be a cause or a useful predictor, X must take place prior to Y. If X happened after Y it is not a cause[3]. Sometimes researchers cannot effectively observe or operationalize time. For example, the moods of roommates are theoretically timed causes of each other but may unfold so quickly that they appear simultaneously causal (Siegel & Alloy, 1990). It is possible that there are nanoseconds in between, but these are unobservable. Furthermore, excessive complexity of timings and multitudinous mood causes running in both directions leave the researcher viewing mood effects as simultaneous.

Macro-comparative research is similar on a larger time scale. Contextual data tend to measure time points spanning an entire year. Reciprocally causal effects that take place in just days, weeks or even months subsume into these yearly observations. For example, public opinion likely causes changes in policymaking on a weekly or monthly basis as policymakers constantly try to meet public preferences. Simultaneously, public opinion changes within minutes or hours in response to policy changes. When capturing these opinion-policy effects with survey data, the two appear to have simultaneous causality within each year unit. Moreover, survey researchers lack yearly comparative opinion data across countries, e.g., cross-sectional yearly time-series[4], rendering longitudinal methods sometimes inappropriate. Having sporadic macro-comparative survey data means SFMs might be appropriate, but this is not a sufficient condition to use them. Theory must drive this decision (Hayduk et al., 2007; Kaplan, Harik, & Hotchkiss, 2001).

Given a theory of simultaneous feedback between two phenomena, I label them Y_1 and Y_2[5], where at least two different linkages exist between them if not more. One for the effect of Y_1 on Y_2 and one vice-versa. However, when I observe and quantify Y_1 and Y_2 as variables, they have only one empirical linkage: their covariance (or correlation). Identifying two effects statistically, when there is only one covariance, is not possible. Y_1 and Y_2 are *nonrecursive* meaning that their respective effects on each other cannot be identified using only their joint information. Their reciprocal relationship makes them *endogenous* meaning caused from

3 The method herein applies to causal or explanatory research subsuming causes or several causes into a package of predictive power without considering the mechanisms in detail. Although causality is at the heart of the theoretical side of SFMs, the vast realm of mathematics and philosophy of causality is beyond the scope of this paper (Pearl, 2010; Sobel, 1996).

4 Although impressive, many macro-comparative sources of survey data barely qualify as longitudinal, cross-sectional time-series when fielded only every 2 to 10 years (e.g., *European Social Survey, World Values Survey and International Social Survey Program*).

5 I use Y_1 and Y_2 rather than X and Y, because Y denotes dependent variables. Reciprocally causal variables are dependent on each other.

within; however, identifying these nonrecursive endogenous effects requires some *exogenous* causes from without.

I describe this problem using Equations 1 and 2, and Figure 1. Both cases present a system logically *underidentified* – there are more parameters to be estimated than pieces of observed information (two coefficients b_1 and b_2 yet only one covariance of Y_1 and Y_2).

$$Y_1 = b_1 Y_2 + e_1 \tag{1}$$

$$Y_2 = b_2 Y_1 + e_2 \tag{2}$$

Regression analysis could estimate Equations 1 and 2, but results are probably inaccurate given a theory of reciprocal causality. In Figure 1 the arrows represent theoretical effects, and b_1 and b_2 represent regression coefficients. Y_1 is not known without knowing Y_2 and Y_2 is not known without knowing Y_1: An endless circle!

Identifying b_1 and b_2 is an exercise in finding more variables or parameters. Figure 2 gives four common formal models containing reciprocal causality, some identified, others not. Adding *instrumental variables* (*IV*s) enables identification of unique b_1 and b_2 effects. An *IV* is *exogenous*: not caused by the system described in the model, not caused by Y_1 or Y_2 and not moderating or somehow causing the causal paths linking Y_1 and Y_2. Figure 2A describes some phenomenon labeled Y_1 occurring at time "t" that is both a cause (arrow pointing away) and outcome (arrow pointing towards) of another phenomenon Y_2 measured at the same time. In this, IV_1 must be a cause of Y_1 but not of Y_2; and IV_2 must cause Y_2 but not Y_1 (see section "Instrumental variables").

Figure 2A is the basic SFM form.

Other common reciprocal effects models appear in Figure 2B-2D. Cross-lagged reciprocal effects (2B) are a common form of reciprocal causal modeling (for discussions: Billings & Wroten, 1978; Schaubroeck, 1990). Looking at Y_1 and Y_2 longitudinally over time generates separate, unique covariances between Y_1 and Y_2; one for $Y_{2,t\text{-}1}$ with $Y_{1,t}$ and another for $Y_{1,t\text{-}1}$ with $Y_{2,t}$. Cross-lagged models require the assumption that Y_1 and Y_2 do not cause each other simultaneously for identification (omitted arrows between them at time t). Macro-comparative survey researchers rarely have sequential time series of survey data in several countries making these models untenable, often because of missing time points or the exact

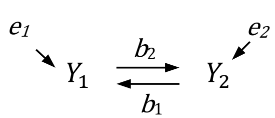


Figure 1 Path Model of Equations 1 and 2

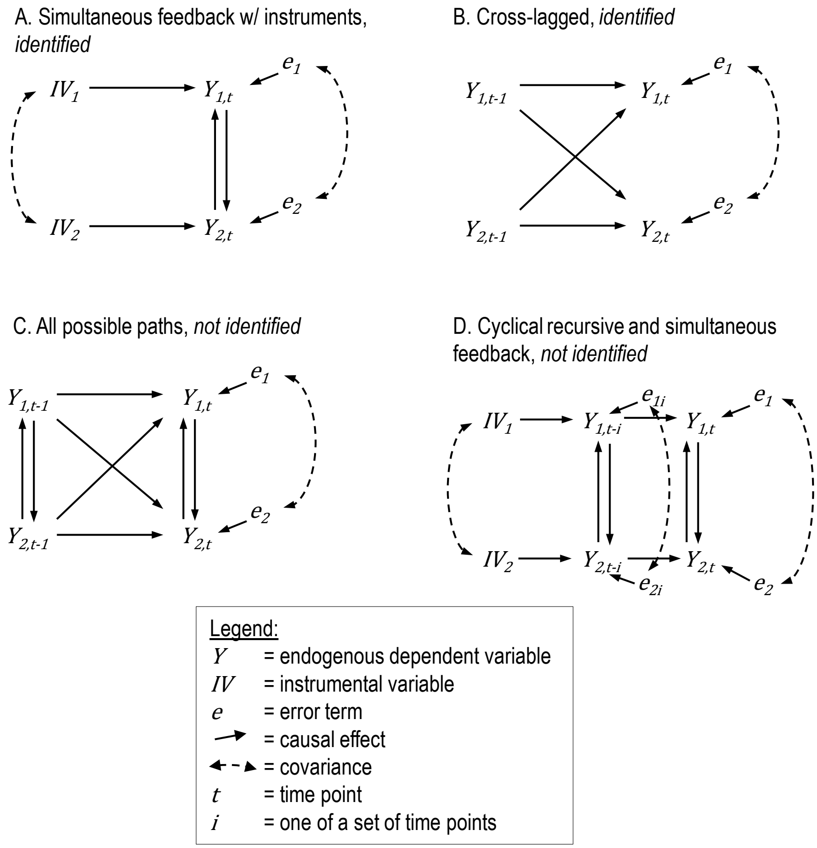


Figure 2 Various Models of Reciprocal Causality

timing of cause and effect do not match the starting and ending points of the survey (Finkel, 1995). If causes occur at a less-than-yearly interval, in addition to across time-units, then Figure 2C is accurate visually but underidentified statistically. A similar story occurs when adding instrumental variables to 2C as shown in 2D. The instruments do not add enough power to overcome the cyclically recursive problem of observing Y_1 and Y_2 over time because they are causes of their later selves in addition to causing each other leading again to too many parameters.

Conditions Necessary for Simultaneous Feedback Models

A strong theory, equilibrium, model identification and appropriate instrumental variables are the necessary features to employ Figure 2A.

Theory

The first and most important requirements of SFMs are theoretical. Without theory, the two arrows connecting Y_1 and Y_2 do not exist. There must be an *a priori* logic to the data-generating model, defensible against confounding effects (Heckman, 2000; Rigdon, 1995). Thus, *a theory of simultaneous causality is the baseline condition*. This theory must specify that during the observational window causal effects materialized between Y_1 and Y_2; regardless of whether these are direct or operating through intermediary mechanisms. A researcher must provide sufficient argument for simultaneity. That of, (1) *co-determinacy* with effects that happen 'instantaneously' in less time than can be observed, or (2) *complexity* with effects that are constantly taking place going in many directions having various lengths of time to complete; so as to appear simultaneous. Without this theoretical basis to the Y_1 and Y_2 relationship, researchers have no ground to stand on in defense of simultaneous feedback (Hayduk et al., 2007; Markus, 2010). Theory determines the design of a formal path model, instrumental variables, equilibrium, size and direction of effects, the set of independent variables, and the nature of errors and estimation techniques. Suffice to say, theory is paramount.

Equilibrium

Two forms of equilibrium need be present in SFMs. The first is that causal effects are theoretically stable or behave in a stable manner. There should be logical argument that the impact of Y_1 on Y_2 and vice-versa, do not change over time (Kaplan et al., 2001). In other words, the effects should not depend on when in time the researcher observes Y_1 and Y_2 (Sobel, 1990). This is a grey area as inevitably all social things change over time; so a better stance to defend might be they do not change much in a given period. For example, if the area of farmed land reduces the hunger in a society while the rate of hunger increases the area of farmed land, a researcher might argue for equilibrium, as a change in one produces a predictable change in the other. Statistically speaking the regression coefficients should be stable. However, technology increases food produced per acre, disrupting the equilibrium because each acre has a larger impact on hunger reduction. This implies that the regression coefficients change if technology changes, but might be stable before and after. If the model includes events before and after this change, it is misspecified as a SFM.

The second part is that the causal effects are part of a context at equilibrium, e.g., a political or judicial system. If a system experiences shocks then equilibrium is unlikely, e.g., disruptive wars or economic recessions. Therefore, the researcher must rule out changes to the larger systems within which Y_1 and Y_2 operate (see section "Disequilibrium").

Identification

Any formal model, including one with simultaneous feedback must be identified to produce meaningful results or results at all[6]. To identify two statistical coefficients that capture two theoretical effects between Y_1 and Y_2 there must be more than one covariance in the model. Only one covariance in the feedback model is *underidentified,* meaning more parameters to estimate than pieces of observed information leading to a negative value for *model degrees of freedom*. Pieces of observed information are all parameters the researcher observes in the data including the means, variances and covariances of the variables in the model, also known as model "elements" (Rigdon, 1994). In SFMs, the observed means are often not estimated because researchers' main interests are in the coefficients between Y_1 and Y_2 that derive entirely from covariances, irrespective of means. Adding means to the analysis generally complicates things with few cases.

Without means, the formula to calculate pieces of observed model information is $v(v+1)/2$, where v is the number of observed variables (Kline, 2011). The model needs a minimum of the same number of model elements as freely estimated parameters for identification, i.e., model degrees of freedom needs to be larger than or equal to zero. To illustrate, I add one predictor variable X, as shown in Figure 3. Figure 3A is not identified because it requires estimation of four coefficients (**a** through **d**) and three variances (**g** through **i**), with residual covariance **m** optional. Fixing **m** to zero for now, and knowing nothing about **a** through **i**, there are seven freely estimated parameters (**a** through **i**). That means I need seven pieces of information for a just-identified model. There are only six pieces in Figure 3A: three covariances ($X, Y_1 \mid X, Y_2 \mid Y_1, Y_2$) and three variances (for $X, Y_1 \,\&\, Y_2$), or 3(4)/2=6. Thus, model degrees of freedom is smaller than zero (six minus seven). Figure 3A is underidentified.

Figure 3B includes IV_1 and IV_2, creating 5(6)/2 = 15 pieces of information. Assuming that the *IV*s and the error terms are correlated (parameters **n** and **m** respectively), the model has 15 freely estimated parameters (all letters in 3B),

6 Any introductory text on structural equation modeling covers identification. I find Kline (2011) a useful source.

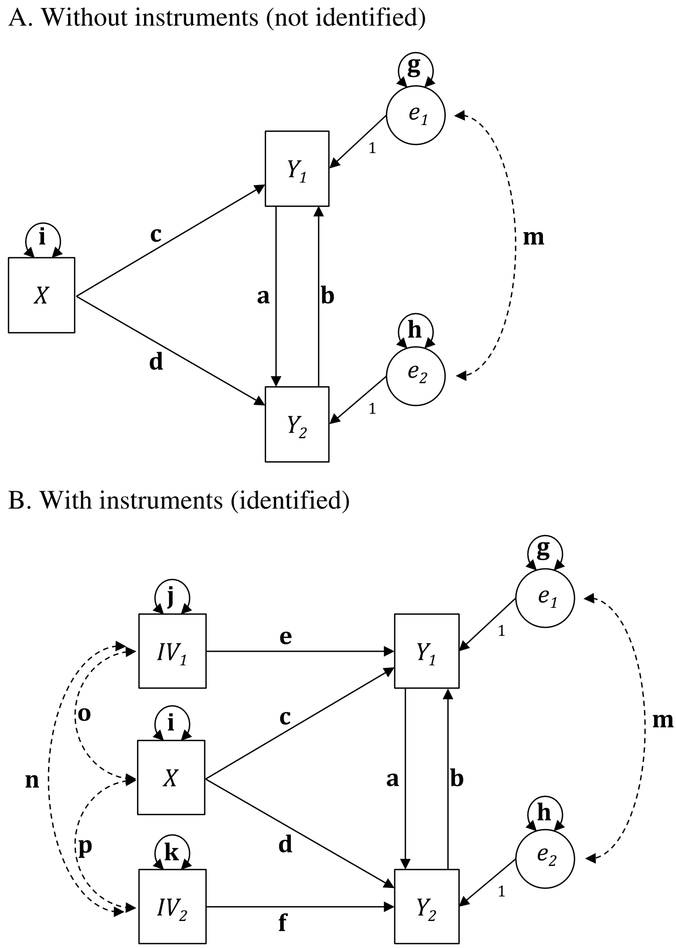


Figure 3 Identifying Simultaneous Feedback Models

meaning model degrees of freedom is zero and the model is just-identified. An ideal model has more than zero, for example three *IV*s leads to 6(7)/2=21 pieces of observed information and 20 freely estimated parameters; degrees of model freedom equals one. However, *IV*s are difficult to find. An identification rule requires at least one *IV* for each *Y* variable. If both instruments are attached to Y_2, and none to Y_1, the model might have degrees of freedom greater than zero, but the model is still not identified without an *IV* for Y_1. This is known as the *rank condition*. This condition is satisfied when, "each variable in a feedback loop has a unique pattern of direct effects on it from variables outside the loop" (Kline, 2011, p. 135). Adding more X variables to 3B does not help with identification as it does not change the degrees of model freedom nor add *unique* direct effects.

Instrumental Variables

Identification depends on instrumental variables (IV_1 and IV_2). Necessary conditions for selecting *IV*s are theoretical and statistical. "Instrumental variables" is both an estimation technique and a label for specific exogenous variables (Sargan, 1958). This section is devoted to exogenous *variables*, saying nothing of estimation techniques[7]. An *IV* must be exogenous to the dependent variable. In experimental language, *IV* causes the distribution of a treatment but not the outcome. In non-experimental language, the endogenous variable depends on the values of the *IV* independently from the dependent variable, or the dependent variable only shows covariance with the *IV* after conditioning on the endogenous variable.

In Figure 3B, the *IV* for Y_1 must not cause Y_2. If IV_1 is a cause of Y_2 then IV_1 is an independent variable, not an *IV*. All independent variables explain or predict all endogenous variables, thus are part of the data-generating model of Y_2 (and Y_1). For IV_1 to pass it must not be part of the data-generating model of Y_2. This is the *exclusion restriction*. The problem is not correlation of IV_1 with Y_2, but correlation of IV_1 with e_2; i.e., correlation with the unexplained disturbance or error in the dependent variable after adjusting for the impact of all independent variables. If IV_1 causes Y_2, or omitted variables cause both IV_1 and Y_2 then a correlation of IV_1 with e_2 exists; and the larger this correlation, the larger the problems with the *IV*. If IV_1 has a small correlation with e_2 because of measurement or random error, then as the sample size approaches infinity the correlation approaches its true value of zero (i.e., asymptotic correlation = 0). If so, small IV_1 with Y_2 correlations after adjusting for covariates are acceptable.

When meeting these conditions, IV_1 and IV_2 decompose the single correlation between Y_1 and Y_2 in Figure 3B into 3 parts: (1) the part that could result from a causal effect or a shared omitted causal effect of Y_1 on Y_2 (covariance left after removing that predicted by IV_2), (2) the same for Y_2 on Y_1, and (3) the unexplained remaining covariance of error terms e_1 and e_2. Although technically optional, Part (3) is usually modeled, because finding instruments that explain *everything* about Y_1 and Y_2 with no remainder is unlikely. Moreover, the error term e_1 is produced by a causal effect of Y_2 (path **b**). Yet e_2 is a part of Y_2 and is therefore *by definition* a part of the error term e_1, i.e., correlated with its own partial correlation produced from Y_2 being regressed on Y_1 (Wong & Law, 1999, p. 73). The same is true for e_2, and therefore specifying no residual correlation may deny the causally defined

7 Other literature covers this in-depth (Angrist, Imbens, & Rubin, 1996; Angrist & Krueger, 2001; Bascle, 2008; Bollen, 2012).

model its own properties. Thus, sometimes a cross-sectional nonrecursive model with correlated errors is the 'best available' approximation of cross-lagged reciprocal effects when they are otherwise underidentified.

Even if theoretically not causal, a large correlation between IV_1 and Y_2 is a problem statistically. The larger the correlation the more variance that all independent variables must explain in Y_2 before IV_1 is left uncorrelated with e_2. In other words, the partial correlation of IV_1 and Y_2 takes away variance in IV_1 that is necessary to explain Y_1. Thus, the larger this correlation, the greater the disruption of the researcher's goal to explain variance in Y_1 independent of Y_2 and all independent variables. An inverse of this problem occurs when IV_1 has an increasingly closer-to-zero correlation with Y_1 (Bartels, 1991). The smaller the correlation, the less unique variance of Y_1 that can be explained by IV_1. These two conditions describe a *weak instrument* problem. Theoretical arguments establish exclusion restrictions necessary to use instrumental variables; however, statistics help identify potential weak instrument problems.

In SEM, model diagnostics, in particular modification indices provide a simple first line of defense to identify weak instruments (see "Fit testing and diagnostics"). This applies because the structural model (what the researcher draws in a path diagram and then programs into the statistical software) fixes the correlation of each *IV* with each corresponding *e* to be zero. The fit and modification indices tell the researcher if these fixed zero correlations are realistic given the data. Alternatively, traditional weak instrument tests come from estimating whether results from the instrumental variable estimator and the OLS estimator are consistent, defined in a number of ways depending on the test (Bollen, 2012; Hahn & Hausman, 2002).

There are a variety of statisticians arguing for statistical methods to identify instrumental variables without theoretical arguments that an *IV* meets the exclusion restriction (see "Other concerns"). Although these methods may asymptotically recover a *known* causal effect (as shown in simulations), the SFM researcher is searching for causal effects whose existence or size is empirically *unknown*. If already known, research becomes unnecessary. Moreover, even when the correlation of IV_1 and Y_2 is exactly zero, there is no statistical way to know for sure that IV_1 and e_2 do not correlate due to causal or omitted variable linkages. Suppression or omitted variables can easily produce a statistical relationship of zero, when the actual causal relationship is non-zero (MacKinnon, Krull, & Lockwood, 2000)[8]. Thus, theoretical arguments are necessary to rule out 'backdoor' or confounding relationships among variables. Finally, arguments must establish that the

8 The drawing of a causal structure with a path diagram or graph notation introduced by Wright (1920) allows researchers to follow rules determining d-separation, exogeneity, collision, and confounding. However, the drawing of the model depends entirely on qualitative use of reason and logic (not statistics or data) (Chen & Pearl, 2015; Elwert, 2013).

instrument is applicable to all cases in the data. If there are cases where the instrument might have a unique causal relationship with the independent variable, so that effects are not *monotonic*, then this is another form of confounding calling for model re-specification.

Although focused on experimental research, a meta-analysis of instrumental variable estimates in political science suggests that researchers routinely fail to offer theoretical arguments that the IV is: (1) unrelated to unobserved/omitted causes of Y, (2) has no direct (causal) effect on Y, and (3) that the instrument could plausibly affect all cases (Sovey & Green, 2011)[9]. This neglect has grave implications for the trustworthiness of results.

An Application – Opinion and Policy

I use the example of Breznau (2017) modeling simultaneous feedback between public opinion and social spending to provide a didactical picture of SFMs. I only briefly summarize the theory from the original research, to keep the focus on execution of the SFM. Public opinion and social policy are an example of theoretical simultaneous feedback, because: (1) Opinion and policy are co-determinant occurring at the same moments or overlapping moments in time. Observing public opinion in a one-year unit prevents observation of anything other than simultaneous effects, even if multiple effects take place within a year. (2) The relationship is so complex that a simultaneous model may come closer to reality than something with arbitrary lags (as taken from years of a survey). Policymakers imagine opinion or act on expected future changes in opinion before opinion changes occur, while public opinion responds to policymakers' intentions and discussions before they actually change policy. Moreover, opinion responds to many things at once over many points in time and the responses take different lengths to materialize. The same applies to policymaking. Given all these effects starting, maturing, declining and then stopping over time, I expect that there is a simultaneous effect, or average simultaneous effect underlying all effects.

The instruments I employ are female labor force participation (IV_1) for public opinion (Y_1) and veto points (IV_2) for policy (Y_2). Labor force participation influences policy attitudes. Holding male participation roughly equal (as seen across OECD countries), variation in the distribution of female participation links to changes in aggregate opinion. Women, who are significantly more supportive of social policy than men are, become less supportive when in the labor force, on average. Moreover, the policy 'styles' of different countries show no patterning by female labor force participation suggesting that at least in recent decades it has no effect on social policy in the aggregate (i.e., exogenous from Y_2). Veto points deter-

9 An argument I am guilty of not making in Breznau (2017)!

mines how easy it is to block legislation in the design of the political system (e.g., executive or minority veto, bicameralism or federalism), thus where veto points are higher, policy provisions should be lower. Veto points are part of a larger institutional framework of societies that might influence public opinion; however, previous research suggests that they are independent (i.e., exogenous from Y_1). Moreover, veto points predate the measurement of public opinion by decades if not centuries, further meeting the exclusion restriction (see Breznau, 2017).

The data I use are publically available; public opinion in the *International Social Survey Program* 'Role of Government' and 'Religion' modules and social policy spending from the *Organization for Economic Co-operation and Development* 'Social Expenditures Database' covering 70 country-time points (across 1985-2006). I provide the variances and covariances necessary to estimate the main models. I include means only for didactic purposes (see Appendix 1-Table A1). All variable measurements and countries are in Appendix 1-Table A2, reproducing Breznau (2017, p. 597). Almost all SEM software reads raw data or covariance matrix data (including correlation/variance matrices). Appendix 1-Table A3 provides programming code (some call this "syntax") for *Mplus*, *Stata* and *R* (*RStudio* running *lavaan*). *Stata* and *R* allow programming the matrix by hand, and *Mplus* reads a .dat file, which is a product of copying the matrix into a text editor and saving it with the file extension .dat[10].

I analyze models of opinion and policy reflecting Figure 3B with four independent *X* variables (aged population, right-party power, unemployment and GDP) predicting both *Y* outcomes. Table 1 presents results for M1, a model of free estimation with little theory and no additional model constraints. Column "b" are unstandardized ('metric') coefficients, and "β" standardized coefficients. The results from *Mplus* here are identical to the other software except rounding error.

The results reveal how much Y_1 and Y_2 cause or explain each other's variance. The standardized coefficient for Y_2 predicting Y_1 suggests that social policy has a *very large* impact on public opinion (0.715), larger than public opinion has on social policy (0.084). However, according to standard testing the effects are insignificant. The insignificance of the smaller effect is perhaps not surprising but insignificance of the very large effect demonstrates the difficulty in disentangling reciprocal effects statistically. Moreover, the countries are not exactly a sample of a larger population, like with human populations. Cut-offs (e.g., $p<0.05$) are perhaps arbitrary without a sample population to generalize into. The t-statistic is still useful for gauging the coefficients. Thus, Y_2 impacting Y_1 is more reliable and precise ($t=0.148/0.088=1.682$) than vice-versa (at 0.357).

10 A1-Appendix One is at the end of this document. The long-form of all code, data, and supplementary analyses are available in Appendix Two and Three, A2 and A3 at https://osf.io/gyz6p/, and .dat files at https://osf.io/cxzj6/.

Table 1 Results from M1. Freely Estimated Simultaneous Feedback between Opinion and Policy

Y_1 (public opinion) ON	b	s.e.	β	Fig 3B label
Y_2 (social policy)	0.148	0.088	0.715	**b**
X_1 (aged)	0.024	0.116	0.052	$\mathbf{c_1}$
X_2 (right)	-0.659	0.656	-0.133	$\mathbf{c_2}$
X_3 (unemp)	-0.070	0.039	-0.264	$\mathbf{c_3}$
X_4 (GDP)	-0.055	0.024	-0.287	$\mathbf{c_4}$
IV_1 (FLP)	-0.073	0.018	-0.540	**e**
Y_2 (social policy) ON				
Y_1 (public opinion)	0.403	1.129	0.084	**a**
X_1 (aged)	1.134	0.318	0.507	$\mathbf{d_1}$
X_2 (right)	-4.615	2.560	-0.194	$\mathbf{d_2}$
X_3 (unemp)	0.187	0.140	0.145	$\mathbf{d_3}$
X_4 (GDP)	0.113	0.140	0.124	$\mathbf{d_4}$
IV_2 (veto)	-7.509	2.988	-0.235	**f**
	variance		*std.variance*	
$e.Y_1$	0.630	0.323	0.654	**g**
$e.Y_2$	13.211	2.242	0.592	**h**
	covariance		*correlation*	
($e.Y_1$,$e.Y_2$)	-1.878	1.213	-0.651	**m**

Note. b are metric and β are standardized coefficients; 70 country-time point cases from ISSP, OECD and other data sources (see A1-Table A2 or Breznau 2017, M10B); Figure 3B contains only one *X* variable so labels include a subscript to differentiate the four *X* variables in this model.

Scholars should exercise caution when interpreting effects independently. The relationship is a loop, not a single causal arrow. Here this loop accounts for (0.715*0.084=0.06) 6% of the joint distribution of the two *Y* variables (although this percentage also depends on the signs and scaling of the coefficients, see section "Explaining variance"). If correctly specified, social policy is a stronger component of this loop. In fact, the term *field* better describes this relationship because the forces are simultaneous and constant like magnets. The coefficients represent constant forces in this stable field. This contrasts with a cyclical loop where a change in one variable sends effects looping through Y_1 and Y_2 in a cyclical process. A steady-state force of the loop and a cyclical force running through the loop are different. To say that the levels of Y_1 on Y_2 are at equilibrium because of their perpetual effects on each other is different than stating that causal effects between Y_1 on Y_2 unfold in specific, precise periods.

I do not rule out the cyclical version of feedback, but have specific theoretical arguments for a non-cyclical version, one that takes place without yearly-time consideration and is sufficiently complex to warrant SFMs. I might take interest in the cyclical relationship when investigating a specific social policy with specific time periods of voting or policymaking. But this macro-comparative exercise presumes that the sum of all specific instances contains common simultaneous feedback; i.e., not particular to one country-year. The comparative advantage here is the ability to test if the general process formulated in a theory of simultaneous feedback and positive returns can be explained by these data (Breznau, 2017; Pierson, 2000).

Without acknowledging reciprocal causality in some form, scholars might measure a unidirectional effect of Y_1 on Y_2 and then separately estimate unidirectional Y_2 on Y_1 rather than a SFM. Appendix 1-Table A4 reveals results from separate regressions. The striking difference is that in both unidirectional regressions the β-coefficients for Y_1 and Y_2 are close to 0.1. This approach leads researchers to conclude that either public opinion explains or causes social policy (Y_1 causes Y_2) or vice-versa (Y_2 causes Y_1), and in either case that the effect is around magnitude of 0.1 standard deviations. Given a theory of simultaneous or reciprocal causality, both conclusions are false and these models are misspecified[11]. The theory used in constructing M1, and the non-zero loop effect of 6% are evidence of this misspecification.

Hypothesis Testing – The SEM Perspective

All parameters in M1 are free, showing how causal effects *might* look if I know nothing theoretically about Y_1 and Y_2 feedback. Given a sufficiently detailed theory of simultaneous feedback, a scholar knows something about the feedback. Thus, I test hypotheses derived from this knowledge. This is the structural equation modeler perspective focusing on overidentified models (Bollen, 1989). This perspective aims to test if a hypothetically derived model leads to something not far off from observational data. If the implied covariances of an overidentified model are not significantly different from observed covariances, then the hypothetical model may reflect the real-world data-generating processes. Testing hypotheses means comparing models with different exclusions or constraints to determine which fits the data better. Both model testing and model comparison require overidentified models.

11 For example, Zhu and Lipsmeyer (2015) use ISSP data to show an impact of policy on opinion while Brooks and Manza (2006) use ISSP data to show an impact of opinion on policy without acknowledging reciprocal causality in their models.

Adding more instrumental variables achieves overidentification, as each adds one degree of model freedom. However, instrumental variables are rare and having two here represents the current limits of this research, beyond speculation (Breznau, 2013, p. 132; 136).

Fixing Parameters

Arguments for a reciprocal relationship of Y_1 and Y_2, are likely to include theory of what this relationship looks like. This is true for opinion and policy feedback (Pierson, 2000; Soroka & Wlezien, 2010). Thus, I specify hypotheses about the nature of the feedback and fix parameters to reflect this. The methodological advantage is an overidentified model. The theoretical advantages are testing competing hypotheses to construct improved theory.

After reviewing the literature I determine that a thermostatic feedback theory suggests that the standardized coefficient **a** (from Figure 3B) is negative 0.05 and **b** is positive 0.30 (see Breznau, 2017). I fix the parameters to these values in M2. The SEM software analyzes only unstandardized effects, thus it is necessary to derive them by scaling the standard deviation of the standardized variable from one to its observed value[12]. Meanwhile an increasing returns theory suggests that both coefficients are positive, possibly around 0.15 as specified in M3. The code is in Appendix 1-Table A5, and Table 2 presents the results.

The other variables' coefficients do not carry much in the way of hypothesis testing (that comes in "Fit testing and diagnostics"); however, they should match theoretical expectations. For example, if the coefficient for aged (X_1) was large and negative, I would become very suspicious that my model is misspecified because it is well-established that more older persons in a society requires far more social spending and usually means greater support of social spending.

A researcher might wish to fix an error term, covariance or mean instead of an effect. M4 has a fixed Y_2 error variance of 0.3, fixed covariance of Y_1 and Y_2 error terms at zero and means of Y_1 and Y_2 at zero. I do not have theoretical arguments for these constraints, they are didactic. Survey data provide the possibility to calculate measurement error for public opinion and I invent the number 0.3 here to represent this possibility. A fixed covariance of zero would be that the model represents a closed system accounting for all possible causal pathways between the variables. This would meet an experimental ideal, where the model explains all things that cause Y_1, Y_2 and the causal loop between them. But this is highly unlikely in the complex realm of cross-national survey research (see section "Instrumental vari-

12 Standardized effect formula: $\beta = b * \frac{\sigma_X}{\sigma_Y}$; metric effect formula: $b = \beta * \frac{\sigma_Y}{\sigma_X}$; where β = standardized coefficient, b = metric coefficient, σ_X = standard deviation of the independent variable, and σ_Y = standard deviation of the dependent variable.

Table 2 Models of Competing Theories of Opinion-Policy Simultaneous Feedback

	M2			M3			M4		
variable	b	s.e	β	b	s.e	β	b	s.e	β
Y_1 (public opinion) ON									
Y_2 (social policy)	-0.010	- -	-0.048	0.030	- -	0.146	0.030	- -	0.165
X_1 (aged)	0.216	0.038	0.466	0.167	0.037	0.362	0.209	0.027	0.484
X_2 (right)	-1.434	0.413	-0.291	-1.240	0.402	-0.252	-1.055	0.331	-0.229
X_3 (unemp)	-0.034	0.028	-0.129	-0.044	0.027	-0.165	-0.006	0.018	-0.023
X_4 (GDP)	-0.053	0.020	-0.281	-0.053	0.019	-0.282	-0.044	0.016	-0.249
IV_1 (FLP)	-0.063	0.015	-0.471	-0.066	0.014	-0.494	-0.045	0.009	-0.358
Y_2 (social policy) ON									
Y_1 (public opinion)	1.500	- -	0.311	0.750	- -	0.154	0.750	- -	0.137
X_1 (aged)	0.901	0.211	0.403	1.062	0.207	0.474	1.175	0.164	0.495
X_2 (right)	-3.376	2.264	-0.142	-4.217	2.225	-0.177	-3.929	2.211	-0.156
X_3 (unemp)	0.172	0.142	0.134	0.183	0.140	0.142	0.245	0.121	0.180
X_4 (GDP)	0.210	0.104	0.229	0.148	0.103	0.160	0.201	0.080	0.206
IV_2 (veto)	-8.070	3.107	-0.252	-8.369	2.986	-0.261	-7.183	2.987	-0.212
$e.Y_1$	0.446	0.075	0.466	0.424	0.072	0.445	0.300	- -	0.360
$e.Y_2$	13.702	2.318	0.613	13.234	2.240	0.589	13.370	2.260	0.532
$(e.Y_1,e.Y_2)$	-0.279	0.307	-0.113	-0.472	0.293	-0.199	0.000	- -	0.000

Note. *Stata* results shown; *R (lavaan)* and *Mplus* identical except rounding error. M4 is not theoretical, has didactic purpose only.

ables”). Nonetheless, I constrain it here for exercise. Means at zero is not important theoretically, it just centers the expected values of Y_1 and Y_2[13].

Fit Testing and Diagnostics

Tests of fit determine how well a theoretically derived model explains real-world observations or compares with alternative models. There is a small universe of these tests. The art of ruling out alternative theoretical models is crucial to scientific utility (Hayduk et al., 2007; and discussed on the structural equation modeling listserv SEMNET), and primarily comes from investigation of how close the

13 Researchers may have a theory that effects **a** and **b** are equal, but not have any prediction about their size. It is possible to constrain **a** and **b** to equality and let computer estimation decide what size is ideal in all three softwares (see A3-Appendix Three).

model-implied covariances come to the freely observed covariances in the data. The proportion of explained variance (r^2) is often a secondary concern. The term *residual* denotes the differences between model-implied covariances and observed covariances. *Residual* also describes OLS error (in $\hat{Y}$), thus structural modelers sometimes use *fitted residuals* or *covariance residuals* to adjudicate these concepts (Kline, 2011).

For just-identified models (like M1) the covariance residuals are zero as implied and observed are identical. In overidentified models, larger residuals suggest worse local fit. Scholars rely on *standardized residuals* and *normalized residuals* given that residuals on their own do not have a common metric. Appendix 1-Table A6 provides residuals for M2 and M3. Smaller residuals support M3.

I might worry about the -1.28 normalized residual of IV_2 and Y_1 in M2 (Appendix 1-Table A6). This suggests unexplained covariance remaining between these variables, where none should be present. This might evidence a weak instrument. However, M3 is the preferred model where this residual is slightly lower at -0.964. Given that M3 fits well overall (as shown in Table 3), and that the theory supports the instrument of veto points being exogenous to public opinion, I tentatively defend IV_2. Yet future research should search for other *IV*s. What causes policy changes that does not cause opinion changes is a puzzle. Finding strong and valid instruments is a perpetual concern (Antonakis et al., 2010).

The *model chi-square* (χ^2) provides the primary statistic for evaluating global model fit. The χ^2 comes from maximum-likelihood estimation (for a good introduction see Kline, 2011, p. 199). The *exact fit hypothesis* is that implied and observed covariance matrices are identical except for random error. Put into test terms, χ^2 difference should not be significant at $p<0.05$, otherwise the matrices in comparison *are* significantly different offering evidence to reject this model. Thus, $p>0.05$ is a reasonable level to not reject the exact fit hypothesis. If this test passes, it does not guarantee the strength of the *IV*, but asserts that nothing about the model radically departs from the observed data; i.e., displays reasonable global fit. The exact fit test becomes increasingly likely to fail the larger the sample because it is more likely to pick up very small confounding parameters in the empirical realm. In macro-comparative survey research, having too large of a country sample is unlikely a problem. The *equal fit hypothesis* is that two implied covariance matrices do not differ from one another. If $p<0.05$ they are significantly different supporting the larger model (with less degrees of freedom). Note that models are only comparable with an equal fit test when they are *nested*; i.e., have all the same basic parameters and observational data.

There are several other global fit diagnostics. Considering all of them is helpful in selecting models, especially when they are not nested (Kline, 2011)[14]. Table

14 David Kenny's website provides discussions of model fit http://davidakenny.net/cm/fit.htm.

Table 3 Model Fit Statistics and Tests

Statistic	Test	Interpretation	Arguments	M1	M2	M3	M4
χ^2				0	5.456	2.758	18.078
df		Model df	0=just identified	0	2	2	6
P-value	Exact fit	Significance of implied and observed covariance differences	p>0.05 equal covariances	*NA*	0.065	0.252	0
P-value[a]	Equal fit (cf. M4)		p<0.05 smaller model is worse	*NA*	0.013	0.004	*NA*
AIC		lower is better	decrease of 5-10 better	1957.1	1958.6	1955.9	1967.2
BIC		lower is better	decrease of 5-10 better	1990.8	1987.8	1985.1	1991.9
RMSEA		lower is better	good <0.05, bad>0.10	0	0.157	0.074	0.170
P-value	P-close	significance of one-sided test RMSEA is greater than 0.05	p>0.05 rejects	*NA*	0.097	0.311	0.016
CFI		closer to 1.0 better	good >0.95	1	0.959	0.991	0.855
TLI		closer to 1.0 better	good >0.95, problem >1.0	1	0.731	0.941	0.686
SRMR		standardized difference of implied and observed correlation residuals	0 = identical	0	0.028	0.020	0.048

Note. AIC "Akaike's Information Criterion", BIC "Bayesian Information Criterion", RMSEA "Root Mean Square Error of Approximation", CFI "Comparative Fit Index", TLI "Tucker-Lewis Index", SRMR "Standardized Root Mean Square Residual"

[a] Equal fit test uses χ^2 and degrees of freedom statistics as the difference of the current model from M4, then a χ^2 table reveals significance levels, or researchers can use an online calculator or Excel command "=CHISQ.DIST.RT(χ^2;df)"

3 contains fit and diagnostics for models M1-M4, offering some preferable targets of these indices. I conclude that M3 is better than M1 because M1 does not have a strong theory to test and AIC and BIC are worse; and better than M2 because all fit indices (AIC, BIC, RMSEA, CFI and TLI) are better. Also, exact fit is less significant (0.252 vs. 0.065) and equal fit more significant (p-value 0.004) than M2 (0.013). It is better than M4, although M4 is just for example.

In addition to residuals, another tool to identify local misfit is *modification indices*. For every parameter in the model, the modification index is the change in χ^2 if that parameter (coefficient or residual covariance) were freely estimated instead of estimated in its current form. The values are zero for parameters already freely estimated and take on positive values for parameters currently fixed (for example the effect of IV_1 on Y_2 in all of the models). Appendix 1-Table A7 lists all non-zero modification indices for M2 and M3. Appendix 1-Table A7 suggests that estimating a free parameter for the regression of Y_2 on IV_1 is a way to improve the model. The normalized residual between Y_2 and IV_1 is -1.28 (see Appendix 1-Table A6) supporting this claim; however, a much larger gain in model fit would result from adding a freely estimated coefficient for Y_1 on IV_2 (4.374 in M2) than for Y_2 on IV_1 (0.745 in M2). This distinction is not evident from looking only at the residuals. Yet, neither of these is possible because the model is not identified with the addition of either parameter (as per the rank condition discussed earlier). Here again are the current limits of this research.

Modification indices are agnostic statistical scores; they do not identify a theoretical problem. Thus, simply freeing parameters in the model might defy, disrupt or debunk the causal model that the researcher carefully constructed using theory. Modification indices are a tool for researchers to use to re-visit their theories and discover what might be missing logically, *before* making any changes to the model. Focusing on M2: In Table A7, the modification indices are identical for the effect of IV_2 on Y_1 and Y_2 on Y_1, and identical for IV_1 on Y_2 and Y_1 on Y_2. This demonstrates how endogeneity works in the SFM. There is residual covariance between Y_1 and Y_2 (normalized value of 0.197 in M2) and the fit of the model may suffer as a result, as the modification index of 4.374 suggests. This essentially means there is a statistical relationship (covariance) between Y_1 and Y_2 not explained by the model and if something could account for this unique feedback error, the model would fit better; in this case a better or additional instrument for IV_2. I did not discuss this in Breznau (2017), but this is a useful finding from this excursus pointing at further research.

Explaining Variance

Sometimes a purpose of explaining variance arises in addition to fit testing. In a SFM, this is a difficult conceptual task. The loop is *the product of both coefficients* (effects **a** and **b** in Figure 3B and Table 1) running between Y_1 and Y_2. In M1, the *loop causal effect* of Y_2 on Y_1 is not 0.715, but includes the effect of Y_1 on Y_2 of 0.084 as an indirect effect, and thus (0.715*0.084)=0.06. To calculate this effect as a percentage, take $1/(1 - Y_1{*}\ Y_2) = 1/(1 - 0.06) = 1.064$ = the original amount plus *6.4%* (Paxton, Hipp, & Marquat-Pyatt, 2011). One cycle through the feedback loop produces about 6.4% of the endogenous variables' covariance[15]. To this loop causal effect we may apply a Sobel-like test revealing a significance score (z-value) of 0.131[16]. Interpretation is identical to a t-test making this statistic non-significant, which is not surprising given that the coefficients are not significant. Normally, another cycle would recover an additional 6% of 6% of the original covariance and so forth. *In SFMs, there is no perpetual looping effect.* One loop is the theoretically specified 'number of cycles' for the SFM (Hayduk, 2009). The ideal model M3 has a loop causal effect of 2.25% (=0.03*0.75), lower than the 6% found in M1, but offering the best theoretical loop causal effect from this research based on fit diagnostics.

The loop causal effect only offers the amount of unique covariance explained by the loop. The remainder may be of interest to the researcher; however, the amount of explained variance of Y_1 and Y_2, like their path coefficients, are reciprocally related[17]. The error of either Y variable actually contains part error and part non-error coming directly from the other endogenous variable's error and thus violating the definition of error in OLS regression. The non-error part is not a component of the theory underlying the model, but an implication of the feedback loop.

Hayduk (2006) proposes a re-specification of r^2 to resolve this problem called the blocked-error-r-square (beR^2). Perfectly appropriate for SFMs, it equals the percentage of variance explained by the model when excluding the other error term as predictor (i.e., the non-error). The beR^2 in M2 is (0.517/0.959=) 0.539 or 53.9% for Y_1 and (9.887/22.366=) 44.2% for Y_2, and for M3 the values are 56.1% for Y_1 and 41.7% for Y_2 (see A3-Appendix Three). The results say little about differences between the models; in fact, they point out that modeling two very different theoret-

15 The formula accounts for situations with opposite signed coefficients, or coefficients greater than one. As in any statistical model, all indirect effects should be calculated from unstandardized coefficients, thus the loop causal effect is (0.148*0.403)=0.06. Although the causal effect should be identical regardless of calculation method, always rely on unstandardized ('metric') coefficients.

16 The standard error (SE) of loop causal effect (where the two causal paths **a** and **b** from Figure 3B are subscripted and normal font "b" is a metric coefficient) is: $SE_{ab} = \sqrt{b_a^2 SE_a^2 + b_b^2 SE_b^2}$; the significance test is then $b_a b_b / SE_{ab}$.

17 Although beyond the scope here, it is interesting to think about the direction of this residual covariance. In infinite looping cycles, a negative covariance approaches zero while a positive covariance explodes towards infinity. In the SFM, there is only one cycle, but there is an implied force of direction suggesting that unobserved causes push away from equilibrium (positive) or towards it (negative).

ical perspectives leads to similar explained variances. Given the small sample-size-to-variables-ratio, it is not surprising that these models explain so much variance.

I did not discuss this in Breznau (2017), that simultaneous feedback accounts for just over 2% of the joint distribution of public opinion and social spending. This would be trivial in standard r-square logic, but this is literally the explained variance unique to the loop itself. The feedback loop is like its own independent variable explaining variance in Y_1 and Y_2. Moreover, this begs the question: what is the loop? It represents the simultaneous impact of public opinion and social policy on one another. This simultaneity occurs in roughly one-year observation windows. Adding more observations should not change this if the loop is stationary at equilibrium. Therefore, disturbances to opinion or policy at best impart a 2% shift in the distribution of opinion and policy. If speaking in terms of majority elections this could make the difference in outcomes. In terms of social spending, this would impart an increase of 60 units (Dollars, Euro, Yen, etc) if a social benefit provides 3,000 units for something (pension, unemployment, etc). These potential outcomes suggest 2% may be non-trivial.

Further Considerations

Estimators

The task of the estimator is to identify what results most closely fit the implied covariance matrix to the observed covariance matrix (Myung, 2003). The most common estimator for this task is maximum likelihood (ML), or one of its many variants. In econometrics instrumental variables estimation often involves two- or three-stage least squares (2SLS or 3SLS) estimators. For SFMs, ML is the least biased estimator because it takes into consideration all information in the system (i.e., both equations) simultaneously. However, misspecification can lead ML to larger bias than 2SLS under some conditions (Paxton et al., 2011). This potential tradeoff suggests that the researcher may gain from running sensitivity checks with 2 or 3SLS to identify misspecification (Kirby & Bollen, 2009), but should not use the results because they are counter to a theory of simultaneity. 2SLS violates the assumption that the errors are correlated (**m** in Figure 3) because it removes the error through instrumental variable stages. However, as noted long ago by economists, any adjustment to one outcome variable or its error term feeds back into the other and estimating the equations separately misses this process (Hausman, 1983, p. 194; Pearl, 2015).

The key is whether unobserved causes (and effects) are randomly distributed with respect to the reciprocally causal relationship of Y_1 and Y_2. If they are not, then the researcher can have little faith in the estimation of **a** and **b** in Figure 3, and should reconsider the formal model rather than worrying about estimators. The default in all three software packages and the default for researchers should be ML.

Disequilibrium

If there are meaningful changes in the size or direction of a causal force during the observation period, then SFMs may not be the appropriate tool. Kaplan, Harik and Hotchkiss (2001) demonstrate some risks associated with estimation under disequilibrium. They simulated different systems that experienced a shock before moving back to equilibrium. They took cross-sections out of the data series to estimate SFMs to test the severity of violating the equilibrium assumption. Their findings reveal that both regression coefficients representing the causal effects between endogenous variables (c.f., Y_1 and Y_2 herein) change somewhat dramatically as the system goes from the shock toward its equilibrium point. The error terms follow a similar pattern. The change in size of coefficients is gradual and smooth in the case of systems that move toward equilibrium without major fluctuations; however, when simulating a system with big oscillations the changes to the regression coefficients are sporadic if not chaotic. In either case, the problem is non-ignorable.

A researcher could mistakenly estimate model Figure 2A when in fact the correct model is 2D wherein $Y_{1,t-i}$ shapes $Y_{1,t-1}$ which leads to a new cycle of effects between Y_1 and Y_2, and then $Y_{1,t-1}$ takes on an entirely new causal effect on $Y_{1,t}$ because of whatever transpired in the first loop (arrows between Y_1 and Y_2) at *t-1*. This means that the model is *cyclically recursive* instead of nonrecursive (Billings & Wroten, 1978). Unfortunately, it is not possible to test for equilibrium, because the data needed for such a test are missing by definition. This leaves a strong burden on the researcher to argue for equilibrium. In the case of macro-comparative survey research, useful arguments may arise based on stable political and cultural systems. For example, the welfare states of Western Europe show a strong degree of stability in their political systems after the 1950s; whereas the Communist states of Eastern Europe broke down and experienced the shock of market transition in the 1990s.

In cross-sectional survey data, there are somewhat random assortments of countries and time-periods available, case-in-point are ISSP data. If the effects and system are truly at equilibrium, then it does not matter what random assortment of country-time-points are in the analysis. All should reveal the same effects. Subdividing the sample, it is possible that the timing of surveys provides a sensitivity test. I demonstrate this by splitting the data into all observations prior to 1998 (Group 1) and all those 1998 and later (Group 2) (see Appendix 1-Table A8 for covariances). I run M2 and M3 separately on the split data. Table A9 reveals that M3 is still preferable to M2 in both groups, and that most effects follow similar patterns between the groups. However, the models do not fit nearly as well as when run on the pooled data – as seen from a few basic fit indices. Nonetheless, the χ^2 p-value from the exact fit tests passes and it appears reasonable that effects are stable over time, for all non-missing years. The very small sample sizes are likely to blame for the troubling other indicators. I compare implied covariance matrices for M3 in Appendix 1-Table A10. Here the main test variables in the model (involving IV_1, IV_2, Y_1 and Y_2) carry similar implied covariances across the two groups. A potential problem is X_4 (unemployment), which switches signs for some of the covariances between the groups. This is evidence that further consideration should be given to this variable in future research to see if it is disrupting the stability of the system. Also, *maybe* there was a slightly different size of effects in Group 1 given the model fits the Group 2 data better; although, much more work is necessary here. This sensitivity analysis does not guarantee stability, and although this procedure is not an established method, it follows the art of structural equation modeling to pay detailed attention to model diagnostics.

Other Concerns

Missing values. Strictly speaking missing values should be dealt with in the estimation of the model as opposed to imputing them separately as if they were observed values. The reason for this is that missing values are subject to special measurement error and ignoring this can produce misleading results. However, contextual-level data are not observations in the strict sense of the word. Values for gross domestic product or level of democracy for example stem from complex calculations whose inputs are not necessarily identical across societies. Researchers at organizations such as the OECD take painstaking efforts to make these values as identical as possible. These values do not represent objective qualities of societies in the way that observed variables such as age or height represent objective features of individuals. Contextual variables are instead more abstract. If they are missing it is best to take the nearest available year. The SFM is not suited for imputing values because of endogeneity.

Aggregation and Comparison. Survey data come from micro-level observations, but macro-comparative researchers aggregate them in some way. Researchers should identify population averages, and then use weights and appropriate measurement models, perhaps performing aggregation in several ways as sensitivity analyses. Monte Carlo simulations suggest that idiosyncratic research practices related to weighting and measurement easily impact results in small-N studies (Breznau, 2016). Furthermore, in order to meaningfully use comparative survey data, all questions need the same cognitive meaning in each socio-cultural context (Davidov et al., 2014). Researchers should establish measurement invariance before using survey data, and correct for measurement error using a measurement model and predicted latent scores that account for differential item functioning when there are three or more scale variables. In this example, previous research suggests measurement invariance of the two ISSP questions (Andreß & Heien, 2001)[18]. Given that there are only two items, the loadings are equal. Thus, a predicted 'factor' is identical in variance with simply taking their mean as I did here.

Estimation without instruments. Several authors suggest estimating *IV* models without observed instrumental variables. Theoretically speaking this violates the exclusion restriction. These methods include estimating a latent or model-implied instrumental variable, or finding a subgroup of the total sample where a researcher can identify a causal instrument (Bollen, Kolenikov, & Bauldry, 2014; Ebbes et al., 2005; Heckman, Urzua, & Vytlacil, 2006; Heckman & Vytlacil, 1999). Suffice to say it is possible but not recommended.

Nonlinear models. If the endogenous variables are non-linear, SFMs are still possible using alternative regression estimation techniques. Simply resorting to linear probability models may introduce new forms of bias (Finch & French, 2015; Terza, Bradford, & Dismuke, 2008)

Conclusion

This excursus shows that data limitations of macro-comparative research are not always a burden. With a theory of sub-yearly causal timing, scholars need not automatically reject cross-sectional survey data as a source for investigating their hypotheses. There are many theoretical forms of reciprocal causality for this. The simultaneous feedback model is only one form. Awareness of this method is not a sufficient condition to use it. Every step in the process of modeling simultaneous feedback must have theoretical argumentation behind it. Theory is a necessary condition for employing a simultaneous feedback model. Without a theory to specify the model, there is no identification of the reciprocal effects and probably no identification of the model. Instrumental variables do not appear through random chance or out of thin air. Perhaps those normally running a bunch of correlations or regressions and then trying to explain the results may learn something from simultaneous feedback modeling, because theory is not 'optional' (Kalter & Kroneberg, 2014).

The impetus for bringing light to this method is the fact that so many macro-comparative phenomena in survey research appear to have reciprocal causality, and the forms of causality are highly complex and unfold in imprecise moments in time. There are well established methods, for example cross-lagged, fixed-effects/random-slope, error correction and vector autoregressive models for fitting longitu-

18 Others find similar questions to be measurement equivalent in the ESS (Roosma, van Oorschot, & Gelissen, 2014)

dinal models. Given the correct research design it is possible to integrate simultaneous feedback in a longitudinal model (Geweke, 1982) like an extension of Figure 2D. Whether or not simultaneous feedback can capture both lagged and instantaneous processes is a theoretical consideration, one limited by available data. The loop causal effect from a SFM may then impact other outcomes (Hayduk, 1987). The loop itself acts as an independent 'variable' or a causal force, a consideration that researchers hopefully take away from this excursus.

There are limitations. Although data derive from individual-level sources, I am not aware of the possibility to model a SFM using multi-level techniques nor individual-level measurement models. Ideally, a measurement model is integrated into a path model for a fully parsimonious structural equation model. This would have a single variable for each survey item and their relationship with the latent scale (here public opinion), and it would have two levels of data analysis. Lacking degrees of freedom prevents the former, and a peculiarity of the SFM prevents the latter. The loop only exists at the aggregate level because there is no individual-level variance in social policy. Moreover, public opinion is by definition a group-level phenomenon, meaning strictly macro-level.

Theories germane to simultaneous feedback come in two broad types and both are debatable, so that researchers should use caution. The first type is where forces act upon each other simultaneously in the real world. The possibility of this is a philosophical argument. Some argue that by definition there are actions and reactions in the world, or that all things are reactions to other things. Meanwhile others argue that it is the interaction of objects and actions at the same point in time that constitute causal effects (Mulaik, 2009). Although this paper takes no philosophical position, researchers working with SFMs are by definition stepping on philosophical ground and tapping into debates that stretch throughout the history of social thought. Thus, awareness of these arguments should help researchers defend themselves against epistemological attacks. The second type suggests that simultaneous causality exists without theoretically simultaneous forces, but can be inferred because the window of observation – usually something around a year in surveys – contains enough bi-directional causal forces between two phenomena that it is logical to treat them as simultaneously causal. This means that even though all these effects may run in different directions and have different sizes, that there is a sum or total effect in their causal loop force that is of theoretical and empirical interest.

Although simultaneity across many countries is an interesting comparative perspective to take and test, researchers more often think of comparative research as looking for differences. As my sensitivity analysis in A3-Appendix Three shows, I can compare two different groups in the data, analogous to a moderation analysis. There are theories that opinion and policy will have different sized effects depending on the institutional context (Wlezien & Soroka, 2012), and this presents an exciting avenue for future implementation of simultaneous feedback in macro-comparative survey data in general and specifically in the opinion-policy case.

References

Andreß, H.-J., & Heien, T. (2001). Four Worlds of Welfare State Attitudes? A Comparison of Germany, Norway and the United States. *European Sociological Review*, *17*(4), 337–356. http://doi.org/10.1093/esr/17.4.337

Angrist, J. D., Imbens, G. W., & Rubin, D. B. (1996). Identification of Causal Effects Using Instrumental Variables. *Journal of the American Statistical Association*, *91*(434), 444–455. http://doi.org/10.2307/2291629

Angrist, J. D., & Krueger, A. B. (2001). Instrumental Variables and the Search for Identification: From Supply and Demand to Natural Experiments. *Journal of Economic Perspectives*, *15*(4), 69–85. http://doi.org/10.3386/w8456

Antonakis, J., Bendahan, S., Jacquart, P., & Lalive, R. (2010). On Making Causal Claims: A Review and Recommendations. *The Leadership Quarterly*, *21*(6), 1086–1120. http://doi.org/10.1016/j.leaqua.2010.10.010

Bartels, L. M. (1991). Instrumental and "Quasi-Instrumental" Variables. *American Journal of Political Science*, *35*(3), 777–800. http://doi.org/10.2307/2111566

Bascle, G. (2008). Controlling for endogeneity with instrumental variables in strategic management research. *Strategic Organization*, *6*(3), 285–327. http://doi.org/10.1177/1476127008094339

Billings, R. S., & Wroten, S. P. (1978). Use of path analysis in industrial/organizational psychology: Criticisms and suggestions. *Journal of Applied Psychology*, *63*(6), 677–688. http://doi.org/10.1037/0021-9010.63.6.677

Bollen, K. A. (1989). *Structural Equations with Latent Variables*. New York, NY: John Wiley & Sons.

Bollen, K. A. (2012). Instrumental Variables in Sociology and the Social Sciences. *Annual Review of Sociology*, *38*(1), 37–72. http://doi.org/10.1146/annurev-soc-081309-150141

Bollen, K. A., Kolenikov, S., & Bauldry, S. (2014). Model-Implied Instrumental Variable—Generalized Method of Moments (MIIV-GMM) Estimators for Latent Variable Models. *Psychometrika*, *79*(1), 20–50. http://doi.org/10.1007/s11336-013-9335-3

Brehm, J., & Rahn, W. (1997). Individual-Level Evidence for the Causes and Consequences of Social Capital. *American Journal of Political Science*, *41*(3), 999–1023. http://doi.org/10.2307/2111684

Breznau, N. (2013). *Public Opinion and Social Policy*. Bremen, Germany: University of Bremen Library. https://elib.suub.uni-bremen.de/peid=D00103291

Breznau, N. (2016). Secondary Observer Effects: Idiosyncratic Errors in Small-N Secondary Data Analysis. *International Journal of Social Research Methodology*, *19*(3), 301–318. http://doi.org/10.1080/13645579.2014.1001221

Breznau, Nate. 2017. Positive Returns and Equilibrium: Simultaneous Feedback Between Public Opinion and Social Policy. *Policy Studies Journal 45*(4), 583–612. http://doi.org/10.111/psj.12171

Brooks, C., & Manza, J. (2006). Social Policy Responsiveness in Developed Democracies. *American Sociological Review*, *71*(3), 474–494.

http://doi.org/10.1177/000312240607100306

Chen, B., & Pearl, J. (2015). *Exogeneity and Robustness*. Technical Report R-449 http://ftp.cs.ucla.edu/pub/stat_ser/r449.pdf.

Chong, A., & Gradstein, M. (2007). Inequality and Institutions. *Review of Economics and Statistics*, *89*(3), 454–465. http://doi.org/10.1162/rest.89.3.454

Claibourn, M. P., & Martin, P. S. (2000). Trusting and Joining? An Empirical Test of the Reciprocal Nature of Social Capital. *Political Behavior*, *22*(4), 267–291. http://doi.org/10.1023/A:1010688913148

Davidov, E., Meuleman, B., Cieciuch, J., Schmidt, P., & Billiet, J. (2014). Measurement Equivalence in Cross-National Research. *Annual Review of Sociology*, *40*(1), 55–75. http://doi.org/10.1146/annurev-soc-071913-043137

Ebbes, P., Wedel, M., Böckenholt, U., & Steerneman, T. (2005). Solving and Testing for Regressor-Error (in)Dependence When no Instrumental Variables are Available: With New Evidence for the Effect of Education on Income. *Quantitative Marketing and Economics*, *3*(4), 365–392. http://doi.org/10.1007/s11129-005-1177-6

Elwert, F. (2013). Graphical Causal Models. In S. L. Morgan (Ed.), *Handbook of Causal Analysis for Social Research* (pp. 245–272). Dordrecht, The Netherlands: Springer Science & Business Media. http://doi.org/10.1007/978-94-007-6094-3 13

Finch, W. H., & French, B. F. (2015). Modeling of Nonrecursive Structural Equation Models With Categorical Indicators. *Structural Equation Modeling: A Multidisciplinary Journal*, *22*(3), 416–428. http://doi.org/10.1080/10705511.2014.937380

Finkel, S. E. (1995). *Causal Analysis with Panel Data* (ebook / Qu). Thousand Oaks: Sage.

Geweke, J. (1982). Measurement of Linear Dependence and Feedback between Multiple Time Series. *Journal of the American Statistical Association*, *77*(378), 304–313. http://doi.org/10.1080/01621459.1982.10477803

Hahn, J., & Hausman, J. (2002). A New Specification Test for the Validity of Instrumental Variables. *Econometrica*, *70*(1), 163–189. http://doi.org/10.1111/1468-0262.00272

Hausman, J. A. (1983). Chapter 7 Specification and estimation of Simultaneous Equation Models. In B. T.-H. of Econometrics (Ed.), (Vol. Volume 1, pp. 391–448). Elsevier. http://doi.org/http://dx.doi.org/10.1016/S1573-4412(83)01011-9

Hayduk, L. A. (1987). *Structural Equation Modeling with LISREL: Essentials and Advances*. Baltimore: Johns Hopkins University Press.

Hayduk, L. A. (2006). Blocked-Error-R2: A Conceptually Improved Definition of the Proportion of Explained Variance in Models Containing Loops or Correlated Residuals. *Quality & Quantity*, *40*, 629–649. http://doi.org/10.1007/s11135-005-1095-4

Hayduk, L. A. (2009). Finite Feedback Cycling in Structural Equation Models. *Structural Equation Modeling: A Multidisciplinary Journal*, *16*(4), 658–675. http://doi.org/10.1080/10705510903206030

Hayduk, L. A., Cummings, G., Boadu, K., Pazderka-Robinson, H., & Boulianne, S. (2007). Testing! testing! One, Two, Three – Testing the Theory in Structural Equation Models! *Personality and Individual Differences*, *42*(5), 841–850. http://doi.org/http://dx.doi.org/10.1016/j.paid.2006.10.001

Heckman, J. J. (2000). Causal Parameters and Policy Analysis in Economics: A Twentieth Century Retrospective. *The Quarterly Journal of Economics*, *115*(1), 45–97. http://doi.org/doi.org/10.1162/003355300554674

Heckman, J. J., Urzua, S., & Vytlacil, E. (2006). Understanding Instrumental Variables in Models with Essential Heterogeneity. *Review of Economics and Statistics*, *88*(3), 389–432. http://doi.org/10.1162/rest.88.3.389

Heckman, J. J., & Vytlacil, E. J. (1999). Local Instrumental Variables and Latent Variable Models for Identifying and Bounding Treatment Effects. *Proceedings of the National Academy of Sciences*, *96*(8), 4730–4734. http://doi.org/10.1073/pnas.96.8.4730

Kalter, F., & Kroneberg, C. (2014). Between Mechanism Talk And Mechanism Cult: New Emphases in Explanatory Sociology And Empirical Research. *KZfSS Kölner Zeitschrift Für Soziologie Und Sozialpsychologie*, *66*(1), 91–115. http://doi.org/10.1007/s11577-014-0272-7

Kaplan, D., Harik, P., & Hotchkiss, L. (2001). Cross-Sectional Estimation of Dynamic Structural Equation Models in Disequilibrium. In R. Cudeck, K. G. Jöreskog, & D. Sörbom (Eds.), *Structural Equation Modeling: Present and Future: a Festschrift in honor of Karl Jöreskog* (p. 315). Lincolnwood, IL: Scientific Software International.

Kirby, J. B., & Bollen, K. A. (2009). Using Instrumental Variable Tests to Evaluate Model Specification in Latent Variable Structural Equation Models. *Sociological Methodology*, *39*(1), 327–355. http://doi.org/10.1111/j.1467-9531.2009.01217.x

Kline, R. B. (2011). *Principles and Practice of Structural Equation Modeling* (Third). New York, NY: Guilford Press.

Liska, A. E., & Reed, M. D. (1985). Ties to Conventional Institutions and Delinquency: Estimating Reciprocal Effects. *American Sociological Review*, *50*(4), 547–560. http://doi.org/10.2307/2095438

MacKinnon, D. P., Krull, J. L., & Lockwood, C. M. (2000). Equivalence of the Mediation, Confounding and Suppression Effect. *Prevention Science*, *1*(4), 173–181. http://doi.org/10.1023/A:1026595011371

Marini, M. M. (1984). Women's Educational Attainment and the Timing of Entry into Parenthood. *American Sociological Review*, *49*(4), 491–511. http://doi.org/10.2307/2095464

Markus, K. A. (2010). Structural Equations and Causal Explanations: Some Challenges for Causal SEM. *Structural Equation Modeling: A Multidisciplinary Journal*, *17*(4), 654–676. http://doi.org/10.1080/10705511.2010.510068

Mulaik, S. A. (2009). *Linear Causal Modeling with Structural Equations*. Boca Raton: CRC Press, Taylor & Francis.

Mulatu, M. S., & Schooler, C. (2002). Causal Connections between Socio-Economic Status and Health: Reciprocal Effects and Mediating Mechanisms. *Journal of Health and Social Behavior*, *43*(1), 22–41.

Myung, I. J. (2003). Tutorial on Maximum Likelihood Estimation. *Journal of Mathematical Psychology*, *47*(1), 90–100. http://doi.org/http://dx.doi.org/10.1016/S0022-2496(02)00028-7

OECD. (2012). *Social Expenditures Database*. accessed 01.10.2012: Organization for Economic Cooperation and Development.

Owens, T. J. (1994). Two Dimensions of Self-Esteem - Reciprocal Effects of Positive Self-Worth and Self-Deprecation on Adolescent Problems. *American Sociological Review*, *59*(310), 391–407.

Page, B. I., & Jones, C. C. (1979). Reciprocal Effects of Policy Preferences, Party Loyalties and the Vote. *The American Political Science Review*, *73*(4), 1071–1089. http://doi.org/10.2307/1953990

Paxton, P., Hipp, J. R., & Marquat-Pyatt, S. (2011). *Nonrecursive Models: Endogeneity, Reciprocal Relationships and Feedback Loops*. London: Sage Publications, Ltd.

Pearl, J. (2010). The Foundations of Causal Inference. *Sociological Methodology*, *40*(1), 75–149. http://doi.org/10.1111/j.1467-9531.2010.01228.x

Pearl, J. (2015). Trygve Haavelmo and the Emergence of Causal Calculus. *Economic Theory*, *31*, 152–179. http://doi.org/doi:10.1017/S0266466614000231

Pedhazur, E. J. (1997). *Multiple Regression in Behavioral Research: Explanation and Prediction*. New York: Thomson Learning.

Pierson, P. (2000). Increasing Returns, Path Dependence, and the Study of Politics. *The American Political Science Review*, *94*(2), 251–267. http://doi.org/10.2307/2586011

Rigdon, E. E. (1994). Calculating Degrees of Freedom for a Structural Equation Model. *Structural Equation Modeling: A Multidisciplinary Journal*, *1*(3), 274–278. http://doi.org/10.1080/10705519409539979

Rigdon, E. E. (1995). A Necessary and Sufficient Identification Rule for Structural Models Estimated in Practice. *Multivariate Behavioral Research*, *30*(3), 359–383. http://doi.org/10.1207/s15327906mbr3003_4

Roosma, F., van Oorschot, W., & Gelissen, J. (2014). The Preferred Role and Perceived Performance of the Welfare State: European Welfare Attitudes from a Multidimensional Perspective. *Social Science Research*, *44C*, 200–210. http://doi.org/10.1016/j.ssresearch.2013.12.005

Sargan, J. D. (1958). The Estimation of Economic Relationships using Instrumental Variables. *Econometrica*, *26*(3), 393–415. http://doi.org/10.2307/1907619

Schaubroeck, J. (1990). Investigating Reciprocal Causation in Organizational Behavior Research. *Journal of Organizational Behavior*, *11*(1), 17–28. http://doi.org/10.1002/job.4030110105

Siegel, S. J., & Alloy, L. B. (1990). Interpersonal perceptions and consequences of depressive-significant other relationships: A naturalistic study of college roommates. *Journal of Abnormal Psychology*, *99*(4), 361–373. http://doi.org/10.1037/0021-843X.99.4.361

Sobel, M. E. (1990). Effect Analysis and Causation in Linear Structural Equation Models. *Psychometrika*, *55*(3), 495–515. http://doi.org/10.1007/bf02294763

Sobel, M. E. (1996). An Introduction to Causal Inference. *Sociological Methods & Research*, *24*(3), 353–379. http://doi.org/10.1177/0049124196024003004

Soroka, S. N., & Wlezien, C. (2010). *Degrees of Democracy: Politics, Public Opinion and Policy*. Cambridge: Cambridge University Press.

Sovey, A. J., & Green, D. P. (2011). Instrumental Variables Estimation in Political Science: A Readers' Guide. *American Journal of Political Science*, *55*(1), 188–200. http://doi.org/10.1111/j.1540-5907.2010.00477.x

Svensson, R., Dahlberg, S., Kumlin, S., & Rothstein, B. (2010). *The QoG Social Policy Dataset*. University of Gothenburg: The Quality of Government Institue.

Terza, J. V, Bradford, W. D., & Dismuke, C. E. (2008). The Use of Linear Instrumental Variables Methods in Health Services Research and Health Economics: A Cautionary Note. *Health Services Research*, *43*(3), 1102–1120. http://doi.org/10.1111/j.1475-6773.2007.00807.x

Thornton, A., Axinn, W. G., & Hill, D. H. (1992). Reciprocal Effects of Religiosity, Cohabitation, and Marriage. *American Journal of Sociology*, *98*(36891), 628–651. http://doi.org/10.1086/230051

Wlezien, C., & Soroka, S. N. (2012). Political Institutions and the Opinion–Policy Link. *West European Politics*, *35*(6), 1407–1432. http://doi.org/10.1080/01402382.2012.713752

Wong, C.-S., & Law, K. S. (1999). Testing Reciprocal Relations by Nonrecursive Structural Equation Models Using Cross-Sectional Data. *Organizational Research Methods*, *2*(1), 69–87. http://doi.org/10.1177/109442819921005

Wright, S. (1920). The Relative Importance of Heredity and Environment in Determining the Piebald Pattern of Guinea-Pigs. *Proceedings of the National Academy of Sciences*, *6*(6), 320–332. http://doi.org/10.1073/pnas.6.6.320

Xie, F., & Levinson, D. (2010). How Streetcars Shaped Suburbanization: A Granger Causality Analysis of Land Use and Transit in the Twin Cities. *Journal of Economic Geography*, *10*(3), 453–470. http://doi.org/10.1093/jeg/lbp031

Zhu, L., & Lipsmeyer, C. S. (2015). Policy Feedback and Economic Risk: The Influence of Privatization on Social Policy Preferences. *Journal of European Public Policy*, *22*(10), 1489–1511. http://doi.org/10.1080/13501763.2015.1031159

Appendix 1: Additional Tables

Table A1 Public Opinion and Social Policy Covariance Structure Data[a]

			Y_1	Y_2	X_1	X_2	X_3	X_4	IV_1	IV_2
	Means		0.085	21.370	14.830	0.150	7.300	25.600	52.000	0.348
	Variance		0.976	22.658	4.537	0.040	13.764	26.936	53.729	0.022
	Standard Deviations		0.988	4.760	2.130	0.200	3.710	5.190	7.330	0.149
	variable	*label*	Y_1	Y_2	X_1	X_2	X_3	X_4	IV_1	IV_2
correlatIons	Public Opinion	Y_1	1.000							
	Social Spending	Y_2	0.348	1.000						
	Aged	X_1	0.413	0.532	1.000					
	Right	X_2	-0.141	-0.193	0.052	1.000				
	Unemp.	X_3	0.294	0.128	0.017	0.004	1.000			
	GDP	X_4	-0.405	0.041	0.082	-0.140	-0.525	1.000		
	FLP	IV_1	-0.527	0.003	-0.030	-0.164	-0.585	0.572	1.000	
	Veto	IV_2	-0.068	-0.199	0.053	-0.013	-0.064	0.175	-0.191	1.000
	variable	*label*	Y_1	Y_2	X_1	X_2	X_3	X_4	IV_2	IV_1
covariances	Public Opinion	Y_1	0.976							
	Social Spending	Y_2	1.638	22.658						
	Aged	X_1	0.869	5.397	4.537					
	Right	X_2	-0.028	-0.184	0.022	0.040				
	Unemp.	X_3	1.077	2.253	0.003	0.130	13.764			
	GDP	X_4	-2.076	1.015	-0.145	0.902	-10.117	26.936		
	FLP	IV_1	-3.818	0.098	-0.241	-0.470	-15.895	21.772	53.729	
	Veto	IV_2	-0.010	-0.141	0.000	0.017	-0.035	0.136	-0.209	0.022

[a] Taken from Breznau (2017).

Table A2 Variable Names and Definitions[a]

Name	Type	Measurement	Source
Public Opinion	Endogenous Dependent Variable	Two-item scale from respondents level of agrrement with the responsibility of government to provide jobs and reduce income differences.	ISSP Role of Government (I,II,III,&IV) and Religion (I&II) modules[b]
Social Spending	Endogenous dependent variable measuring Social Policy	The amount of spending on social policy provisions, mostly pensions, employment, unemployment, and health care expressed as a percentage of GDP in the same year.	OECD (2012); also known as "SOCX"
Aged	Independent variable	Percent of the population over age 64.	OECD Social Indicators Data
Right	Independent variable	Percent of national government seats held by right parties.	Svennson et al. (2012); Quality of Government Data
Unemp.	Independent variable	Percent of the labor force that is unemployed.	OECD Social Indicators Data
GDP	Independent variable	Gross Domestic Product at PPP.	OECD Social Indicators Data
Female LFP	Instrument for Public Opinion	Percent of the total female population in the labor force.	OECD Social Indicators Data
Veto Points	Instrument for both Social Policy variables	A scale of institutional measures for the amount of chances a policy has to be vetoed. Based on the work of Lijphart (1999).	Svennson et al. (2012); Quality of Government Data

a This Table is copied from Table 1 in Breznau (2017). See original article for full citations. All variables are measured simultaneously at the current year of the endogenous variables.

b Country-time points are: Australia ('86,'90,'93,'97,'98,'07), Austria ('86,'93,'98), Canada ('96,'00,'06), Denmark ('98,'08), Finland ('06), France ('97,'98,'06), Germany ('86,'90,'91,'96,'98,'06), Ireland ('91,'96,'98,'06), Italy ('86,'90,'96,'99), Japan ('96,'98,'06), the Netherlands ('91,'98,'06), New Zealand ('91,'97,'98,'06), Norway ('90,'91,'96,'98,'06), Portugal ('99,'06), Spain ('96,'98,'07), Sweden ('96,'98,'06), Switzerland ('98,'99,'07), Great Britain ('86,'90,'91,'96,'98,'06), the United States ('86,'90,'91,'96,'98,'06).

Table A3 Programming Code for Simultaneous Feedback Models

Mplus	*R, lavaan*	*Stata*

Data:
FILE IS /data/pospC.dat;
!The file pospC.dat must be specified above
TYPE IS CORRELATION STDEVIATIONS;
!also accepts ‚MEANS‘ and ‚COVARIANCE‘
NOBSERVATIONS ARE 70;
Variable:
NAMES ARE
y1 y2 x1 x2 x3 x4 iv1 iv2;
	library(lavaan)
	#means in same order as cov matrix
	m.mean <- ‚
	0.085 21.37 14.83 0.15 7.3 25.6 52 0.348‘
	#line breaks can go anywhere in the list
	cov <- ‚0.9761 1.6366 22.6576 0.8691 5.3938
	4.5369 -0.0279 -0.1837 0.0222 0.04 1.0777
	2.2604 0.1343 0.003 13.7641 -2.0767 1.0129
	0.9065 -0.1453 -10.1088 26.9361 -3.8166
	0.1047 -0.4684 -0.2404 -15.9087 21.7604
	53.7289 -0.01 -0.1411 0.0168 -0.0004
	-0.0354 0.1353 -0.2086 0.0222‘
	posp.cov <-
	getCov(cov, names = c(„y1“, „y2“, „x1“, „x2“,
	„x3“,
	„x4“, „iv1“, „iv2“)
		clear all
		ssd init y1 y2 x1 x2 x3 x4 iv1 iv2
		ssd set obs 70
		ssd set means 0.085 21.37 14.83 0.15 7.3 25.6 52 0.348
		ssd set sd 0.988 4.76 2.13 .2 3.71 5.19 7.33 0.149
		ssd set cor 1\ 0.348 1\ 0.413 0.532 1\ -0.141 -0.193
		0.052 1\ 0.294 0.128 0.017 0.004 1\ -0.405 0.041
		0.082 -0.140 -0.525 1\ -0.527 0.003 -0.030 -0.164
		-0.585 0.572 1\ -0.068 -0.199 0.053 -0.013 -0.064
		0.175 -0.191 1
		*use „ssd set cov“ for covariances

Analysis:
TYPE = GENERAL;
!for raw data add the following two lines
!MODEL = NOMEANSTRUCTURE;
!INFORMATION = EXPECTED;
 	
Model:
y1 ON y2 x1 x2 x3 x4 iv1;
y2 ON y1 x1 x2 x3 x4 iv2;
y1 WITH y2;
 	
Output:
STDYX; !include standardized estimates
	m1.model <- ‚
	#Regressions
	y1 ~ y2 + x1 + x2 + x3 + x4 + iv1
	y2 ~ y1 + x1 + x2 + x3 + x4 + iv2
	#Correlated Residuals
	y1 ~~ y2‘
	
	#this saves the results as an object named „fit“
	fit <- sem(m1.model, sample.cov = posp.cov,
	sample.nobs = 70, meanstructure = FALSE)
	#lavaan command to display results
	summary(fit)
	standardizedSolution(fit)
		sem (y2 x1 x2 x3 x4 iv1 -> y1) (y1 x1 x2 x3 x4 iv2 -> y2),
		cov(e.y1*e.y2) nomeans standardized
		*remove „standardized“ for metric estimates

Note. Programmed using *Mplus* 7, *R* (lavaan) 0.5-22 and *Stata* 14; “!” in *Mplus*, “#” in *R*, and “*” in *Stata* are comments; *R* (lavaan) only reads in covariance data (as of 25.03.2017), thus four decimal places used to make estimates as close as possible to the *Mplus* and *Stata* correlation and standard deviation data; variables labeled y1 (social spending), y2 (public opinion), x1 (aged), x2 (right), x3 (gdp), x4 (unemp), iv1 (female labor force participation), and iv2 (veto points).

Table A4 Results from Separate Unidirectional Regressions

Y_1 (public opinion) ON	b	s.e.	β
Y_2 (social policy)	0.029	0.020	0.141
X_1 (aged)	1.789	0.475	0.362
X_2 (right)	-0.117	0.039	-0.252
X_3 (GDP)	-0.044	0.028	-0.165
X_4 (unemp)	-0.053	0.019	-0.280
IV_1 (FLP)	-0.067	0.015	-0.495
var($e.Y_1$)	0.423	0.072	0.441
Y_2 (social policy) ON			
Y_1 (public opinion)	0.559	0.581	0.116
X_1 (aged)	11.721	2.573	0.492
X_2 (right)	-0.417	0.217	-0.187
X_3 (GDP)	0.184	0.140	0.144
X_4 (unemp)	0.127	0.114	0.138
IV_2 (veto)	-7.494	2.986	-0.235
var($e.Y_2$)	13.197	2.231	0.591

Table A5 Code for Fixing Parameters in the Opinion-Policy Feedback Example (see Tables 2 and 3)

<table>
<tr><th></th><th>Mplus</th><th>R, lavaan</th><th>Stata</th></tr>
<tr><td></td><td colspan="3">* * * The upper part of the code for each software is identical to Table 2. * * *</td></tr>
<tr><td>Model Two (M2)</td><td>Analysis:
TYPE = GENERAL;
Model:
y1 ON y2@-0.01;
y1 ON x1 x2 x3 x4 iv1;
y2 ON y1@1.5;
y2 ON x1 x2 x3 x4 iv2;
y1 WITH y2;</td><td>m2.model <- ,

#Regressions
y1 ~ -0.01*y2 + x1 + x2 + x3 + x4 + iv1
y2 ~ 1.5*y1 + x1 + x2 + x3 + x4 + iv2
#Correlated Residuals
y1 ~~ y2‘

fit2 <- sem(m2.model, sample.cov = posp.cov,
sample.nobs = 70, meanstructure =
FALSE)</td><td>sem (y2@-0.01 x1 x2 x3 x4 iv1 -> y1)
(y1@1.5 x1 x2 x3 x4 iv2 -> y2),
cov(e.y1*e.y2) nomeans standardized</td></tr>
<tr><td>Model Three (M3)</td><td>Model:
y1 ON y2@0.03;
y1 ON x1 x2 x3 x4 iv1;
y2 ON y1@0.75;
y2 ON x1 x2 x3 x4 iv2;
y1 WITH y2;</td><td>m3.model <- ,

#Regressions
y1 ~ 0.03*y2 + x1 + x2 + x3 + x4 + iv1
y2 ~ 0.75*y1 + x1 + x2 + x3 + x4 + iv2
#Correlated Residuals
y1 ~~ y2‘

fit3 <- sem(m3.model, sample.cov = posp.cov,
sample.nobs = 70, meanstructure =
FALSE)</td><td>sem (y2@0.03 x1 x2 x3 x4 iv1 -> y1)
(y1@0.75 x1 x2 x3 x4 iv2 -> y2),
cov(e.y1*e.y2) nomeans standardized</td></tr>
</table>

	Mplus	*R, lavaan*	*Stata*
Model Four (M4)	!Change command to include data with means FILE IS /data/pospCM.dat; TYPE IS CORRELATION MEANS STDE-VIATIONS; Model: y1 (e1); !labels the y1 error term [y1@0]; !fix mean/intercept to 0 [y2@0]; y1 ON y2@0.03; y1 ON x1 x2 x3 x4 iv1; y2 ON y1@0.75; y2 ON x1 x2 x3 x4 iv2; !remove correlated y1 WITH y2 error, defaults to 0 Model Constraint: e1 = 0.3; !This fixes the e1 variance to 0.3	m4.model <- , #Regressions y1 ~ 0.03*y2 + x1 + x2 + x3 + x4 + iv1 y2 ~ 0.75*y1 + x1 + x2 + x3 + x4 + iv2 #remove correlated errors y1 ~~ y2, defaults to 0 fit4 <- sem(m4.model, sample.cov = posp.cov, sample.nobs = 70, meanstructure = FALSE)	*variances and means/intercepts fixed as options sem (y2@0.03 x1 x2 x3 x4 iv1 _cons@0 -> y1) (y1@0.75 x1 x2 x3 x4 iv2 _cons@0 -> y2), var(e.y1@0.3) standardized *Remove correlated e.y1 e.y2 errors, defaults to 0

Model Fit	Output:
	STDYX !include standardized results
	SAMPSTAT !observed covariances
	RESIDUAL !covariance residuals
	MODINDICES(0); !modification indices, min 0
		#“object“ must be replaced with model name
		summary(object)
		standardizedSolution(object)
		residuals(object) #raw covariance residuals
		residuals(object, type = „standardized“)
		residuals(object, type = „normalized“)
		modindices(object)
		fitMeasures(object)
			*remove „standardized“ option to get metric results
			estat residuals, norm standardized
			estat mindices, min(0)
			estat gof, stats(all)

Note. M2 and M3 correspond to M11B and M12B in Breznau (2017), and M4 is for didactic purposes on fixing parameters other than coefficients.

Table A6 Structural Residuals for M2 and M3. Observed minus Implied

M2 (Model Two)

Covariance Residuals

var	Y_1	Y_2	X_1	X_2	X_3	X_4	IV_1	IV_2
Y_1	0.004							
Y_2	0.068	-0.032						
X_1	0.000	0.000	0.000					
X_2	0.000	0.000	0.000	0.000				
X_3	0.000	0.000	0.000	0.000	0.000			
X_4	0.000	0.000	0.000	0.000	0.000	0.000		
IV_1	-0.057	1.815	0.000	0.000	0.000	0.000	0.000	
IV_2	-0.022	-0.019	0.000	0.000	0.000	0.000	0.000	0.000

Standardized Residuals

var	Y_1	Y_2	X_1	X_2	X_3	X_4	IV_1	IV_2
Y_1	0.218							
Y_2	0.409	-0.085						
X_1	0.000	0.000	0.000					
X_2	0.000	999	0.000	0.000				
X_3	0.000	0.000	0.000	0.000	0.000			
X_4	0.000	0.000	0.000	0.000	0.000	0.000		
IV_1	-0.497	0.870	0.000	0.000	0.000	0.000	0.000	
IV_2	-2.089	-1.595	0.000	0.000	0.000	0.000	0.000	0.000

M3 (Model Three)

label	Y_1	Y_2	X_1	X_2	X_3	X_4	IV_1	IV_2
Y_1	0.009							
Y_2	0.115	-0.148						
X_1	0.000	0.000	0.000					
X_2	0.000	0.000	0.000	0.000				
X_3	0.000	0.000	0.000	0.000	0.000			
X_4	0.000	0.000	0.000	0.000	0.000	0.000		
IV_1	-0.002	0.395	0.000	0.000	0.000	0.000	0.000	
IV_2	-0.017	0.006	0.000	0.000	0.000	0.000	0.000	0.000

label	Y_1	Y_2	X_1	X_2	X_3	X_4	IV_1	IV_2
Y_1	0.392							
Y_2	0.601	999						
X_1	0.000	999	0.000					
X_2	0.000	999	0.000	0.000				
X_3	0.000	999	0.000	0.000	0.000			
X_4	0.000	999	0.000	0.000	0.000	0.000		
IV_1	-0.027	0.185	0.000	0.000	0.000	0.000	0.000	
IV_2	-1.590	999	0.000	0.000	0.000	0.000	0.000	0.000

Normalized Residuals

M2 (Model Two)									M3 (Model Three)								
var	Y_1	Y_2	X_1	X_2	X_3	X_4	IV_1	IV_2	*label*	Y_1	Y_2	X_1	X_2	X_3	X_4	IV_1	IV_2
Y_1	0.022								Y_1	0.055							
Y_2	0.115	-0.008							Y_2	0.197	-0.039						
X_1	0.000	0.000	0.000						X_1	0.000	0.000	0.000					
X_2	0.000	0.000	0.000	0.000					X_2	0.000	0.000	0.000	0.000				
X_3	0.000	0.000	0.000	0.000	0.000				X_3	0.000	0.000	0.000	0.000	0.000			
X_4	0.000	0.000	0.000	0.000	0.000	0.000			X_4	0.000	0.000	0.000	0.000	0.000	0.000		
IV_1	-0.059	0.442	0.000	0.000	0.000	0.000	0.000		IV_1	-0.002	0.096	0.000	0.000	0.000	0.000	0.000	
IV_2	-1.282	-0.227	0.000	0.000	0.000	0.000	0.000	0.000	IV_2	-0.964	0.074	0.000	0.000	0.000	0.000	0.000	0.000

Table A7 Non-Zero Modificaiton Indices

Freed parameter			M2	M3
Y_1	ON	Y_2	4.374	2.609
Y_1	ON	IV_2	4.374	2.609
Y_1	ON	Y_1	4.374[a]	2.609[a]
Y_2	ON	Y_1	0.745	0.034
Y_2	ON	IV_1	0.745	0.034
Y_2	ON	Y_2	0.745[a]	0.034[a]

Note. "ON" refers to regression coefficients

[a] Variable regression on itself is a statistical artifact of having structural equations (see text).

Table A8 Covariance Structure for Samples Split by Timea

			Group 1 (<1998)[b]								*Group 2 (1998+)*[b]							
			Y_1	Y_2	X_1	X_2	X_3	X_4	IV_1	IV_2	Y_1	Y_2	X_1	X_2	X_3	X_4	IV_1	IV_2
	Means		0.011	21.035	14.189	0.198	8.658	23.262	49.377	0.373	0.059	21.101	15.061	0.125	6.421	27.923	54.221	0.343
	Variance		1.058	26.764	3.921	0.045	21.605	14.544	58.611	0.022	0.957	21.993	5.300	0.033	10.128	30.453	36.418	0.025
	Standard Deviations		1.029	5.173	1.980	0.212	4.648	3.814	7.656	0.147	0.978	4.690	2.302	0.181	3.182	5.518	6.035	0.158
	variable	*label*	Y_1	Y_2	X_1	X_2	X_3	X_4	IV_1	IV_2	Y_1	Y_2	X_1	X_2	X_3	X_4	IV_1	IV_2
correlations	Public Opinion	Y_1	1.000								1.000							
	Social Spending	Y_2	0.563	1.000							0.284	1.000						
	Aged	X_1	0.548	0.554	1.000						0.333	0.540	1.000					
	Right	X_2	-0.169	-0.127	0.005	1.000					-0.109	-0.178	0.203	1.000				
	Unemp.	X_3	0.387	0.148	0.014	-0.200	1.000				0.388	0.311	0.233	0.073	1.000			
	GDP	X_4	-0.340	-0.083	0.245	-0.079	-0.622	1.000			-0.481	-0.120	-0.193	-0.016	-0.530	1.000		
	FLP	IV_1	-0.657	-0.073	0.018	0.086	-0.521	0.604	1.000		-0.558	-0.059	-0.253	-0.281	-0.667	0.529	1.000	
	Veto	IV_2	-0.180	-0.203	0.094	-0.228	-0.269	0.475	0.072	1.000	-0.046	-0.092	0.196	0.175	0.045	0.314	-0.089	1.000

		Group 1 (<1998)[b]								Group 2 (1998+)[b]							
		Y_1	Y_2	X_1	X_2	X_3	X_4	IV_1	IV_2	Y_1	Y_2	X_1	X_2	X_3	X_4	IV_1	IV_2
Means		0.011	21.035	14.189	0.198	8.658	23.262	49.377	0.373	0.059	21.101	15.061	0.125	6.421	27.923	54.221	0.343
Variance		1.058	26.764	3.921	0.045	21.605	14.544	58.611	0.022	0.957	21.993	5.300	0.033	10.128	30.453	36.418	0.025
Standard Deviations		1.029	5.173	1.980	0.212	4.648	3.814	7.656	0.147	0.978	4.690	2.302	0.181	3.182	5.518	6.035	0.158
variable	*label*	Y_1	Y_2	X_1	X_2	X_3	X_4	IV_2	IV_1	Y_1	Y_2	X_1	X_2	X_3	X_4	IV_2	IV_1
covariances: Public Opinion	Y_1	1.058								0.957							
Social Spending	Y_2	2.998	26.764							1.304	21.993						
Aged	X_1	1.115	5.676	3.921						0.751	5.833	5.300					
Right	X_2	-0.037	-0.140	0.002	0.045					-0.019	-0.151	0.085	0.033				
Unemp.	X_3	1.849	3.566	0.131	-0.197	21.605				1.206	4.638	1.703	0.042	10.128			
GDP	X_4	-1.332	-1.640	1.849	-0.064	-11.024	14.544			-2.598	-3.109	-2.458	-0.016	-9.306	30.453		
FLP	IV_1	-5.173	-2.899	0.278	0.140	-18.552	17.644	58.611		-3.295	-1.667	-3.519	-0.307	-12.804	17.622	36.418	
Veto	IV_2	-0.027	-0.154	0.027	-0.007	-0.183	0.266	0.081	0.022	-0.007	-0.068	0.071	0.005	0.022	0.273	-0.085	0.025

[a]Split by ISSP wave. Group 1: Role of Government (1986, 1990 & 1996), Religion (1991); Group 2: Role of Government (2006), Religion (1998).
[b]See Table 2 for variable coding and country time-points.

Table A9 Testing Equilibrium Comparing Results by Group

	M2						M3					
	Group 1 (< 1998)			Group 2 (1998 +)			Group 1 (< 1998)			Group 2 (1998 +)		
var	b	s.e	β	b	s.e	β	b	s.e	β	b	s.e	β
Y_1 ON												
Y_2	-0.010	- -	-0.049	-0.010	- -	-0.050	0.030	- -	0.150	0.030	- -	0.146
X_1	0.330	0.047	0.632	0.116	0.055	0.278	0.275	0.040	0.534	0.069	0.054	0.162
X_2	-0.767	0.435	-0.157	-1.675	0.722	-0.315	-0.681	0.374	-0.142	-1.416	0.707	-0.262
X_3	-0.016	0.026	-0.073	-0.027	0.054	-0.089	-0.021	0.022	-0.098	-0.051	0.053	-0.166
X_4	-0.052	0.035	-0.193	-0.039	0.027	-0.226	-0.057	0.031	-0.215	-0.041	0.027	-0.234
IV_1	-0.079	0.014	-0.585	-0.081	0.031	-0.504	-0.076	0.013	-0.569	-0.089	0.030	-0.549
Y_2 ON												
Y_1	1.500	- -	0.306	1.500	- -	0.302	0.750	- -	0.150	0.750	- -	0.155
X_1	0.895	0.335	0.350	1.122	0.277	0.539	1.154	0.341	0.448	1.201	0.266	0.583
X_2	-1.960	3.213	-0.082	-5.746	3.369	-0.217	-3.206	3.276	-0.133	-6.519	3.244	-0.249
X_3	0.138	0.187	0.126	0.425	0.228	0.282	0.137	0.190	0.125	0.455	0.219	0.305
X_4	0.368	0.246	0.277	0.318	0.140	0.366	0.325	0.251	0.243	0.265	0.134	0.309
IV_1	-7.442	5.128	-0.216	-8.314	4.211	-0.274	-10.944	5.291	-0.315	-7.897	4.043	-0.263
$e.Y_1$	0.253	0.060	0.243	0.494	0.118	0.547	0.187	0.045	0.185	0.476	0.114	0.512
$e.Y_2$	13.250	3.231	0.532	11.875	2.839	0.532	13.752	3.290	0.544	11.008	2.632	0.504
RMSEA		0.261			0.170			0.177			0.080	
CFI		0.934			0.943			0.969			0.987	
Exact p		0.034			0.134			0.123			0.295	

Table A10 Implied Covariance Matricies for M3 by Group

				Group 1 (< 1998)				
var	Y_1	Y_2	X_1	X_2	X_3	X_4	IV_1	IV_2
Y_1	1.009							
Y_2	2.610	25.275						
X_1	1.085	5.512	3.808					
X_2	-0.036	-0.135	0.002	0.044				
X_3	1.798	3.457	0.125	-0.191	20.987			
X_4	-1.296	1.591	1.797	-0.062	-10.711	14.131		
IV_1	-4.981	-1.630	0.265	0.136	-18.010	17.133	56.940	
IV_2	-0.009	-0.124	0.027	-0.007	-0.179	0.259	0.079	0.021
				Group 2 (1998 +)				
var	Y_1	Y_2	X_1	X_2	X_3	X_4	IV_1	IV_2
Y_1	0.928							
Y_2	1.590	21.831						
X_1	0.728	5.768	5.148					
X_2	-0.019	-0.147	0.082	0.032				
X_3	1.173	4.509	1.658	0.041	9.836			
X_4	-2.522	-3.017	-2.382	-0.016	-9.040	29.578		
IV_1	-3.188	-5.018	-3.414	-0.298	-12.443	17.113	35.381	
IV_2	-0.009	-0.066	0.069	0.005	0.022	0.266	-0.082	0.024

Appendix Two and Three

Appendix Two and Three, A2 and A3 available at https://osf.io/gyz6p

methods, data, analyses | Vol. 12(2), 2018, pp. 309-334 DOI: 10.12758/mda.2017.12

Blaming the Young Misses the Point: Re-assessing Young People's Political Participation over Time Using the 'Identity-equivalence Procedure'

Christian Schnaudt[1] *& Michael Weinhardt*[2]

[1] *GESIS - Leibniz Institute for the Social Sciences*
[2] *Bielefeld University*

Abstract

One of the central and constantly recurring features of youth participation studies is the depiction of young people and adolescents as the future of democratic politics. According to previous research, however, young people exhibit generally lower levels of political participation than adults and show decreasing trends in their political activities over time. In this study, we argue that, in order to arrive at meaningful conclusions about young and adult people's political participation over time, 'construct-equivalent' rather than identical instruments of political participation across different age groups and time points should be used. Applying the so called 'identity-equivalence procedure' for political participation across three different age groups and the time period 2002-2014 using data from the European Social Survey (ESS), our results indicate that (1) the concrete manifestations of the concept of political participation differ across young and adult people and over time and (2) levels of political participation are quite similar for young and adult people and increasing from 2002-2014. Therefore, the commonly employed strategy of applying identical instruments of political participation across age groups and time points appears at least questionable.

Keywords: political participation, youth participation, democracy, measurement equivalence, scale development, Mokken scale analysis, European Social Survey

Citizens' participation and engagement in the political process count as a 'conditio sine qua non' of any democratic system. Consequently, it is not surprising that virtually every discussion about the well-being of democracy is strongly linked to debates and complaints about citizens' disengagement and alienation from politics (cf. Verba & Nie, 1972, 1; Verba, Schlozman, & Brady, 1995, 1). In this context, especially young people and adolescents have been singled-out as one of the major driving forces behind decreasing participation rates and growing disenchantment with the political sphere. Common depictions and characterizations of young people and adolescents in previous youth participation studies thus regularly include labels and terms such as 'apathetic', 'alienated', and 'disengaged' (cf. Garcia Albacete, 2011, 2; Martin, 2012, 213). This is especially true in the German context where previous youth participation studies have repeatedly highlighted continuously high levels of political apathy ('Politikverdrossenheit') among the German youth (cf. Schneekloth, 2015, 178-82; Sloam, 2014, 664). As Henn and Foard summarize, "the message from many such studies is that young people's levels of political participation in general are in decline, and at a somewhat more rapid rate than is the case for older adults and also for previous youth cohorts" (2014, 361).

Yet, the validity of such a far-reaching conclusion hinges on several factors, as it implies a simultaneous statement about the levels of political participation (1) for young and adult people as well as (2) over the course of time. In order to allow for this kind of conclusion, a study has to meet at least three criteria. First, it should be based on a coherent sample of *both young and adult people* to facilitate direct comparisons of political participation levels across different age groups. Studies that rely on different samples for young and adult people remain inconclusive as to whether possible differences in political participation levels between age groups are 'real' or merely an artefact of different sampling frames or survey techniques for young and adult people. Second, the sample of both young and adult people should be coherent *over time* to facilitate direct comparisons of participation *trends* across age groups. Third, the *measurement of political participation* should be a valid and reliable representation of the same underlying concept across young and adult people *as well as* over time. This at least necessitates an investigation of the underlying structure of the concept of political participation and at best implies the development of so called 'construct-equivalent' instruments of political participation (cf. Garcia Albacete, 2011, 17) across different age groups and points in time. Studies that simply assume that identical instruments of political participation can be uniformly applied across young and adult people as well as over time without checking this assumption empirically might miss important differences in the

Direct correspondence to
Christian Schnaudt, GESIS - Leibniz Institute for the Social Sciences, Post Box 122155, 68072 Mannheim, Germany
E-Mail: christian.schnaudt@gesis.org

underlying structure of political participation and are therefore ill-suited to draw valid conclusions with regard to young people's political participation.

A cursory glance at existing studies dealing with youth political participation reveals that none fulfills all of the three criteria mentioned. Pure youth participation studies by definition violate the first criterion of a direct comparison between young and adult people (see, for example, Henn & Foard, 2014; Gaiser, de Rijke & Spannring, 2010; Quintelier, 2007). Those studies that conform to the first criterion either lack a comparison over time or do not analyze the underlying structure of political participation (see, for example, Martin, 2012). Finally, those studies that meet the third criterion and assess the underlying structure of political participation are either restricted to one point in time or only investigate one age group, thus violating the first or second criterion (see, for example, Bakker & de Vreese, 2011; Quintelier, 2008; Henn & Foard, 2014).

Against this background, the present study offers a re-assessment of young people's political participation by investigating the structure and levels of political participation across young, adult and old people in Germany over the period 2002-2014. Applying the so called 'identity-equivalence procedure' (cf. Przeworski & Teune, 1966), this study develops age-group and time-point equivalent instruments of political participation that allow for meaningful comparisons of political participation levels across young and adult people as well as over time. In doing so, this study sheds more light on contemporary questions of (increasing) political apathy among young people and the peculiarities of youth participation in general.

The remainder is structured as follows. Section 2 provides an overview of the main findings and strategies of previous studies and identifies common problems in research on political participation in general and youth participation studies in particular. Section 3 introduces the 'identity-equivalence procedure' for investigating the structure and levels of political participation across young and adult people over time. Section 4 illustrates the methods and data used. Section 5 presents the results of the empirical analysis. Section 6 discusses the most important findings as well as their broader implications and concludes.

Research on Political Participation Across Young and Adult People: Main Findings, Strategies, and Problems

One of the central and constantly recurring features of youth participation studies is the depiction of young people and adolescents as hope and sorrow for the future of democratic politics. As Mycock and Tonge (2012, 141) summarize this view, young people are "often discussed within the context of national decline or regeneration, being projected as symbolic of the positive and progressive future or typified as a threat and somehow out of control." Most of the time, however, it is the latter perspective that seems to dominate the discussion. Young people are portrayed as "apathetic or even as antipolitical, with neither aptitude nor inclination for participating in any form of collective social endeavour, and with no sense of civic responsibility" (Henn & Foard, 2014, 360; see also Quintelier, 2007, 165; Neufeind, Jiranek & Wehner, 2014, 278; Martin, 2012, 213; Cammaerts et al., 2014, 648).

Empirically, such negative portrayals are often countered by the observation that young people, while possibly being alienated from traditional electoral or formal politics, do engage in non-electoral or informal modes of political participation that reach beyond the realm of institutionalized politics (cf. Vissers & Stolle 2014, 937; Cammaerts et al 2014, 657; Sloam 2014, 676). In comparison with adults, then, young people's political participation seems to be less formal, less institutionalized, and less hierarchical, and they appear to prefer more individualized, lifestyle-oriented modes of participation such as signing petitions, protesting, or political consumerism (cf. Sloam 2013, 837; Stolle, Hooghe & Micheletti 2005, 250). If these assertions are correct, it is clear that a comparison of political participation between young and adult people does not only have to consider the level of participation, but also the respective modes of participation being used by young and adult people, respectively.

As such, the analysis of young people's political participation is directly linked to discussions about the continuous expansion of the political participation repertoire and distinctions between different 'types' of political participation (cf. van Deth, 2014; Vissers & Stolle 2014, 937). Whereas contacting politicians or working for a political party are usually considered to be specimens of 'formal', 'conventional', 'institutionalized' or 'elite-directed' participation, other modes such as signing petitions, demonstrating or boycotting are usually labeled 'unconventional', 'non-institutionalized' or 'protest' participation (cf. van Deth, 2014, 361; Linssen et al., 2014, 33-4; Marien, Hooghe, & Quintelier, 2010, 198). While such distinctions between different 'types' of participation are well-known and frequently employed in research on political participation, there are at least two problems concerning the way in which they are being used.

The first problem refers to research on political participation in general and touches upon the fact that many studies do not test which of the several modes of participation might actually be summarized to form one (or more) coherent type(s) of political participation. Instead of investigating the structure of different modes of political participation, a lot of studies simply choose to build additive indices (cf. Quintelier, 2007, 174; Hao, Wen, & George, 2014, 1226; Wray-Lake & Hart, 2012, 457) or use self-defined assignments of participation modes to types (cf. Gaiser, de Rijke & Spannring, 2010, 440; Martin, 2012, 218-9; Neufeind, Jiranek, & Wehner, 2014, 285; Soler-i-Marti & Ferrer-Fons, 2015, 101). As a consequence, one and the same mode of participation is oftentimes assigned to different types of participation across different studies. For example, whereas Marien, Hooghe and Quintelier (2010, 198) consider 'donating money' to be a specimen of 'non-institutionalized' participation, Gaiser, de Rijke and Spannring (2010, 440) depict it as a mode of 'conventional' participation. Similarly, Neufeind, Jiranek and Wehner (2014, 285) classify 'signing a petition' as a mode of 'conventional' participation, whereas Gaiser, de Rijke and Spannring (2010, 440) label it as 'unconventional', Martin (2012, 217) as 'non-electoral', and Marien, Hooghe and Quintelier (2010, 198) as 'non-institutionalized' participation. As these examples make clear, previous studies do not assign individual modes to commonly employed types of political participation in a coherent manner. These inconsistencies do not only hamper a comparison of participation levels and trends across different studies, but also leave open the question of whether and which different modes can actually be summarized to form one or more coherent types of political participation.

Those studies that do investigate the structure of political participation provide valuable (empirical) information on which modes form a coherent type of participation, but are usually restricted to one point in time (cf. Bakker & de Vreese, 2011, 457-8; Quintelier, 2008, 359-60; Henn & Foard, 2014, 365). Consequently, these studies have nothing to say about possible changes in the underlying structure of (different types of) political participation over time which, however, is of crucial importance especially in the context of longitudinal studies (e.g., a previously unconventional mode becomes rather conventional over time; see also Linssen et al., 2014; 34).

The second problem, which is more pertinent to our focus on young people's political participation, has to do with the applicability or generalizability of commonly employed conceptualizations and types of political participation across different age groups. Distinctions between different types of political participation, such as 'conventional vs unconventional' or 'institutionalized vs non-institutionalized', belong to the standard toolkit of political participation researchers. The fact that these distinctions are so frequently applied is probably one of the major reasons why their usage is generally not called into question. However, especially in the context of research on youth participation, it appears important to note that these conceptualizations and distinctions have been developed primarily with reference to the general or adult population, which at least leaves room for the possibility that they are not applicable in the same manner to young people as well. As O'Toole et al. remind us, "[y]oung people are often seen in conventional accounts of political participation as simply a subset of the general population. Analyses of youth participation need to consider young people as a specific group with their own particular circumstances and concerns" (2003, 46). In this connection, Quintelier has argued that "young people operate with a very narrow conception of politics that is restricted to formal politics only" (2007, 177; see also O'Toole et al., 2003, 52). If we consider this limited and narrow conception of politics to inform their conception and understanding of political participation as well, young people's political participation may be less facetted and based on fewer modes of participation than that of adult people. In a similar manner, changes or delays in youth transition periods as highlighted by previous studies (cf. Soler-i-Marti & Ferrer-Fons, 2015, 96; Garcia Albacete, 2011, 6) might also lead to varying structures of young people's political participation over time. An empirical investigation of the underlying structure of political participation across age groups *and* over time therefore becomes indispensable in order to shed more light on the differences and similarities concerning the structure, levels and developments of young and adult people's political participation.

The 'Identity-equivalence Procedure' for Political Participation

For our empirical investigation, we make use of the so called 'identity-equivalence procedure' which has originally been introduced by Przeworski and Teune, (1966) in the context of cross-cultural research. The basic premise of this procedure is that, in order to be comparable, measurements of the same concepts do not have to be identical but rather equivalent (cf. Przeworski & Teune, 1966, 555-9). More specifically, as its name suggests, the procedure is based on two consecutive steps. In a first step, it involves the search for a so-called 'identity set' of survey items that can be regarded as a valid representation of a given concept across all subgroups of interest (cf. van Deth, 1986, 265). These subgroups are usually different countries but the same underlying logic can be easily extended to include different social classes or age groups as well. For example, in the present study we search for a common set of survey items that form a consistent scale of the concept 'political participation' across young and adult people alike as well as over time. This common set of items constitutes our 'identity set' of political participation. In a second step, the 'identity-equivalence procedure' implies the search for additional survey items that can be used to extend the identity set of political participation in a subgroup and time-point specific way. Accordingly, in the present study we search – separately for young and adult people as well as time points – for additional survey items that can be added to the existing scale of political participation that is based on the identity set only. Since the respective survey items to be added to the identity scale of political participation possibly differ between young and adult people and time points, the resulting age-group and time-point specific scales of political participation are no longer identical but rather equivalent. Adding age-group and time-point specific items to our identity scale helps us to arrive at "longer, more reliable and more contextually relevant instruments" of political participation (Garcia Albacete, 2011, 29). With this strategy, the 'identity-equivalence procedure' ensures that we are analyzing the same underlying concept across different subgroups and time points (due to the identity scale which consists of the same items across all subgroups and time points) while at the same time allowing for the possibility that manifestations of the same underlying concept might differ in specific ways for different subgroups and time points (due to the construction of the equivalence scales). As such, construct equivalence is achieved by directly building the equivalence scales on the identity scale: "By referring the equivalent indicators back to the identical indicators, this procedure introduces safeguards of validity – the guarantee that the phenomena examined […] constitute specific occurrences of a more general concept" (Przeworski & Teune, 1966, 568).

While the 'identity-equivalence procedure' has been developed for establishing equivalent measures across different cultural contexts, we believe that it can be fruitfully applied to investigate the underlying meaning and structure of the concept political participation across different age groups and time points as well. In contrast to previous studies on youth political participation, we thus do not simply assume that political participation exhibits the same underlying meaning and structure (over time) for young and adult people alike but rather put this proposition to an empirical test.

Methods and Data

For the implementation of the procedure, we rely on Mokken Scale Analysis (MSA) (Mokken, 1971). MSA is based on principles of nonparametric item response theory (IRT) and constitutes a probabilistic extension of the Guttman scale (cf. van Schuur, 2003, 139). MSA can be used to investigate response patterns to a set of survey items that are supposed to measure a certain latent trait, such as 'political participation' in the present study (cf. Sijtsma & Molenaar, 2002; van Schuur, 2003; van der Ark 2007; 2012; Linssen et al. 2014, 39-41; Schnaudt, Walter, & Popa, 2016, 76). MSA assumes that each respondent has a certain, unknown value on that latent trait, so that the probability of a positive response to any of the survey items for political participation increases with that unknown value on the latent trait. For the construction of political participation scales, the individual survey items have to meet certain criteria as implied by the monotone homogeneity model: all item pairs have to be positively correlated and the scalability coefficients for each individual item have to exceed a certain lower bound (usually item H>0.3). In addition, the overall degree of scalability for the resulting scale(s) as indicated by Scale H should exhibit a minimum value of 0.3 as well. In MSA, the item scalability coefficients can be compared to discrimination parameters in parametric IRT models, whereas the Scale H indicates the average discrimination power with regard to the ordering of all items in the final scales (cf. Mokken, 1971, 184-5; van der Ark, 2007, 3-4; Sijtsma, Meijer, & van der Ark, 2011, 33). If the assumptions of the monotone homogeneity model hold, respondents and items can be meaningfully ordered along a latent continuum of political participation.[1] While MSA has been successfully applied in previous studies of political participation (cf. van Deth, 1986; Garcia Albacete, 2011; Linssen et al., 2014), this study is the first to use it for analyzing the structure of political participation across different age groups and time points.

MSA is particularly suitable because it allows us to identify which concrete modes of participation might be summarized to form coherent scales or types of political participation and whether these modes are constant or varying across young and adult people and over time (cf. van Deth, 1986, 265). What is more, it gives us information on the ranking or 'difficulty' of individual survey items along the latent continuum 'political participation' and whether we find an identical or varying item order across young and adult people and over time (cf. Linssen et al., 2014, 42-4; Garcia Albacete, 2011, 24). Finally, it allows us to construct equivalent scales of political participation across young and adult people and over time and thus enables us to draw meaningful conclusions about differences and similarities with respect to the levels of political participation across different age groups and time points.

With regard to our empirical analysis, we rely on German data from the first seven waves of the European Social Survey (ESS) covering the years 2002-2014. The ESS is a biennial survey covering a wide range of European citizens' economic, moral, social and political attitudes and behaviors and has been conducted in more than thirty European countries since 2002 (for a general overview, see Schnaudt et al., 2014). Considering the focus of the present study, the advantage of using data from the ESS consists in its combination of providing (1) a stable set of survey items tapping the concept political participation for a period of twelve years and (2) a representative sample of the German population aged 15 and above. Relying on ESS data thus remedies at least two possible shortcomings of previous studies. First, since it covers the general population aged 15 and above, it enables us to directly analyze differences and similarities in political participation between young and adult people using only one coherent sample. Such a direct comparison between young and adult people allows us to find out more about the specificities of young people's political participation and establishes an advantage vis-à-vis pure youth studies (for example, Gaiser, de Rijke, & Spannring, 2010; Quintelier, 2007). Second, covering people already from the age of 15, the ESS allows us to depict a more realistic and encompassing picture of young people than previous studies relying on a sample only with respondents aged 18 or above (for example, Henn &

1 For a more detailed discussion of MSA, including its properties and underlying assumptions, see Mokken, 1971; Sijtsma & Molenaar, 2002; van Schuur, 2003; van der Ark, 2007; 2012; Ligtvoet et al., 2010, 2011.

Foard, 2014; Wray-Lake & Hart, 2012). Germany is a substantively interesting case to focus on given previous findings about continuously high levels of political apathy ('Politikverdrossenheit') among the German youth (cf. Schneekloth, 2015, 178-82; Sloam, 2014, 664). In addition, our focus on Germany also reflects a pragmatic decision based on sample size and data availability. While the ESS is a survey of the general population, sample sizes in Germany are sufficiently high (more than 2,750 respondents in each of the seven waves) to still allow for meaningful analyses across young and adult people as well as individual waves of the survey (cf. Schnaudt et al., 2014, 501-2). Our focus on Germany thus remedies the problem of very small sample sizes for the young population that is routinely encountered in other studies (cf. Sloam, 2014, 668).

In our following analysis, we employ a total of seven items that are supposed to measure the concept of political participation which we broadly define here as "citizens' activities affecting politics" (van Deth, 2014, 351). While the ESS provides a higher number of suitable items in certain waves, we select these seven items because they are available in all seven waves of the ESS and can be meaningfully applied to all respondents aged 15 and above. This implies that we exclude the item 'voting in national elections' from our analyses as it would lead to the exclusion of a substantial and theoretically important subset of our sample, namely all young people who did not have the chance to vote in the last general election due to their young age (cf. Quintelier, 2007, 169). The seven items selected are: (1) working for a political party or action group, (2) contacting politicians or government officials, (3) working for another organization or association, (4) wearing a badge or campaign sticker, (5) signing a petition, (6) taking part in a lawful demonstration, and (7) boycotting products. The ESS asks which of these several activities respondents have done within the last twelve months.[2] This question wording ensures that responses are not biased against young people who, due to their lower age, did not have the same chances of engaging in political activities as adult people (cf. Martin, 2012, 215). In the remainder of this section, we analyze the structure of these seven items separately for three age groups. In addition to a group of young people (aged 15-29) and a group of adult people (aged 30-65), we also investigate a group of older people (aged 66 and above). This classification is informed by one of the most established findings in participation research according to which political participation follows the shape of an inverted U, implying that participation rates increase with age and then drop again when people get older and reach retirement (cf. Milbrath, 1965, 134). While the cutting point for distinguishing between the second and the third age group is rather straightforward (i.e., transition to retirement), the decision to classify people until the age of 29 as belonging to the young-

2 Each of the seven items is binary in nature (1=have done/0=haven't done). Respondents with missing information ('don't know', no answer, or refusal) on any of the items have been excluded from the analysis (less than 0.5% for each item).

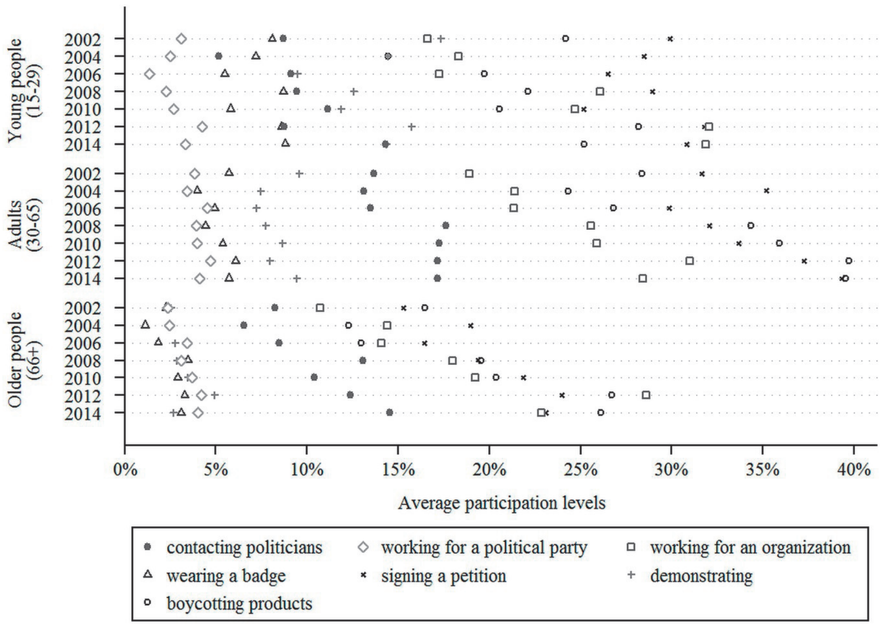
est category follows theoretical arguments and empirical findings about a longer or delayed transition from youth to adulthood (cf. Garcia Albacete, 2011, 6). As "transformations in patterns of youth participation in Western Europe may arise more from the lengthening of youth than from any generational change" (Soler-i-Marti & Ferrer-Fons, 2015, 96), implying that young people reach important stages of their life-cycles (e.g., marriage, getting children) at a later point in time as compared to some decades ago, we consider the age of 29 as a plausible cutting point for distinguishing between young and adult people.

Empirical Findings

Before we turn to the results of the 'identity-equivalence procedure' for political participation, Figure 1 gives a first descriptive overview of the seven items for political participation across the three age groups and seven time points (2002-2014) under consideration.

On average, participation levels are lowest for the oldest age group and highest for the group of adults. The group of young people occupies the medium position. What is more, the figures indicate an increase in the average participation rates for certain items over time across all age groups (e.g., working for an organization, signing a petition, boycotting products). Other forms of participation, such as working for a political party or wearing a badge, remain at rather stable levels across time and age groups.

Having a closer look at the participation profiles of each of the three age groups over time, Figure 1 shows that for young people working for a political party, wearing a badge, and contacting politicians are the least common modes of participation across all years and usually do not exceed participation levels of ten percent. The remaining four items for demonstrating, working for an organization, boycotting products, and signing a petition reach average levels between ten and thirty-two percent across all years but show more variability with regard to their rank order across time. Overall, the participation profile for the youngest age group thus exhibits some internal changes and a certain degree of volatility over time. For the group of adults, a different picture emerges. Here, the general participation profile is very stable over time and exhibits only one minor change with regard to the rank order of the items for signing a petition and boycotting between the years 2006 and 2008. Otherwise, the identical rank order of participation modes is evident across all years. The least common participation modes are working for a political party, wearing a badge, and demonstrating, usually not exceeding average levels of ten percent. The most common modes of participation are boycotting, signing a petition, and working for an organization, with average levels between twenty and forty percent across all years. Contacting politicians occupies an intermediate



Notes: ESS data 2002-2014, data weighted using post-stratification weights.

Figure 1 Average levels of different modes of political participation across three age groups and seven time points (percentages)

position with average levels between thirteen and eighteen percent across the seven time points. Finally, the participation profile of the oldest age group shows the most fluctuations with regard to the rank order of participation modes over time. While wearing a badge, working for a political party, and demonstrating are the least common modes with average levels below five percent, their relative order changes from year to year. The same volatility in the rank order over time holds true for the most common modes of signing a petition, boycotting, and working for an organization, whose levels in all years range between ten and thirty percent. Contacting politicians is the only consistent mode of participation occupying an intermediate rank across all years with levels between seven and fourteen percent.

To summarize, the inspection of the seven individual modes of political participation as depicted in Figure 1 shows some similarities and common trends between age groups and over time. Yet, some differences with regard to the average levels and rank order of these seven modes across age groups and time points are also evident. The main question of interest concerns whether these differences in the frequency distribution and rank order of the seven individual modes indicate the existence of different meanings or structures underlying the concept of political participation across different age groups and time points.

To answer this question, we turn to the 'identity-equivalence procedure' as briefly described before. In a first step, we search for the so called 'identity set' of political participation. The identity set is that set of items which corresponds to the properties of a Mokken Scale and is valid across all age groups and time points under investigation. Starting first with the pooled data set to get an impression of the structure of political participation across all respondents and time points (with no distinctions between age groups and ESS waves), MSA yields a uni-dimensional scale of political participation consisting of six out of the seven items under consideration. More specifically, with the exception of 'boycotting products' all remaining modes of participation can be summarized to form a coherent scale of political participation (Scale H=0.35, LCRC=0.66).[3] This finding also indicates that, at least for the pooled data set, commonly employed types of political participation, such as 'institutionalized vs non-institutionalized', do not receive empirical support. The interesting question at this point is whether the political participation scale found for the pooled data set can be replicated in the same way across all age groups and over time to form our 'identity set' of political participation. The short and clear answer is 'no'. From the seven items included in our analysis, the only set of items that corresponds with the criteria of a Mokken Scale across all age groups and time points consists of the three items working for a political party, contacting politicians, and working for another organization. Accordingly, these three modes of participation can be meaningfully summarized to form our 'identity set' of political participation. Again, it has to be noted that MSA yields only one scale of political participation, indicating that commonly used conceptions and distinctions between different types of political participation are not supported in our data. Table 1 presents the detailed properties of the final three-item identity scale of political participation across age groups and time points.

All item scalability coefficients exceed the critical lower bound of 0.3. The overall scalability of the resulting scales ranges between 0.35 (young people in 2004) to 0.55 (older people in 2012). In five out of twenty-one cases, the scale H is below 0.4 (indicating a weak scale), in eleven out of twenty-one cases the scale H is between 0.4 and 0.5 (indicating a medium scale), and in five out of twenty-one cases the scale H is above 0.5 (indicating a strong scale) (cf. Mokken, 1971, 185). The reliability coefficients of the resulting identity scales as measured by rho and LCRC, respectively, do not reach conventional levels of 0.7, which can be explained by the fact that the identity scale consists of only a small number of three items which, in addition, also lack a uniform distribution in their difficulties (cf. Garcia Albacete, 2011, 27). Lastly, the rank order of the three items within the identity scale is the same across all age groups and time points: The most difficult item is working for a political party, followed by contacting politicians and working for

3 The LCRC (Latent Class Reliability Coefficient) is a measure of reliability in MSA (see van der Ark, van der Palm, & Sijtsma, 2011).

Table 1 Properties of the three-item identity scale of political participation across three age groups and seven time points (item frequencies and scalability coefficients; scale coefficients and reliability)

		Working for political party		Contacting politicians		Working for organisation			
Age group	Year	Item diff.	Item H	Item diff.	Item H	Item diff.	Item H	Scale H	Scale reliability
Young people (15-29)	2002 (N= 525)	.03 (.17)	.50 (.10)	.09 (.28)	.42 (.08)	.17 (.37)	.43 (.08)	.44 (.07)	.52 / .49
	2004 (N= 549)	.03 (.16)	.39 (.13)	.05 (.22)	.35 (.09)	.18 (.39)	.32 (.10)	.35 (.09)	.39 / .34
	2006 (N= 535)	.01 (.12)	.42 (.13)	.09 (.29)	.36 (.08)	.17 (.38)	.42 (.08)	.39 (.07)	.43 / .46
	2008 (N= 457)	.02 (.15)	.52 (.08)	.09 (.29)	.49 (.08)	.26 (.44)	.49 (.09)	.49 (.07)	.49 / .45
	2010 (N= 620)	.03 (.16)	.42 (.11)	.11 (.32)	.36 (.07)	.25 (.43)	.34 (.08)	.36 (.07)	.42 / .38
	2012 (N= 583)	.04 (.20)	.54 (.09)	.09 (.28)	.37 (.08)	.32 (.47)	.44 (.09)	.44 (.08)	.41 / .36
	2014 (N= 531)	.03 (.17)	.45 (.10)	.14 (.35)	.36 (.07)	.31 (.46)	.36 (.08)	.38 (.07)	.42 / .39
Adults (30-65)	2002 (N= 1,849)	.04 (.19)	.58 (.05)	.14 (.34)	.44 (.03)	.19 (.39)	.42 (.03)	.46 (.03)	.58 / .55
	2004 (N= 1,760)	.03 (.18)	.68 (.05)	.13 (.34)	.49 (.03)	.21 (.41)	.48 (.04)	.52 (.03)	.59 / .55
	2006 (N= 1,746)	.04 (.21)	.62 (.05)	.13 (.34)	.48 (.03)	.21 (.41)	.46 (.04)	.50 (.03)	.59 / .56
	2008 (N= 1,696)	.04 (.19)	.58 (.06)	.18 (.38)	.45 (.03)	.26 (.44)	.44 (.03)	.47 (.03)	.56 / .54
	2010 (N= 1,777)	.04 (.20)	.70 (.05)	.17 (.38)	.51 (.03)	.26 (.44)	.52 (.03)	.54 (.03)	.60 / .57
	2012 (N= 1,734)	.05 (.22)	.65 (.05)	.18 (.39)	.50 (.03)	.32 (.47)	.52 (.04)	.54 (.03)	.56 / .53
	2014 (N= 1,822)	.04 (.19)	.59 (.06)	.16 (.37)	.48 (.03)	.27 (.44)	.44 (.04)	.48 (.03)	.56 / .54

Table 1 continued

		Working for political party		Contacting politicians		Working for organisation			
Age group	Year	Item diff.	Item H	Item diff.	Item H	Item diff.	Item H	Scale H	Scale reliability
Older people (66+)	2002 (N= 515)	.02 (.15)	.44 (.10)	.08 (.28)	.38 (.07)	.11 (.31)	.37 (.08)	.39 (.07)	.51 / .49
	2004 (N= 480)	.02 (.15)	.54 (.11)	.07 (.25)	.50 (.07)	.14 (.35)	.36 (.09)	.46 (.08)	.55 / .56
	2006 (N= 567)	.03 (.18)	.66 (.08)	.08 (.28)	.43 (.07)	.14 (.35)	.39 (.07)	.47 (.07)	.57 / .54
	2008 (N= 558)	.03 (.17)	.65 (.09)	.13 (.34)	.44 (.06)	.18 (.38)	.43 (.06)	.47 (.06)	.58 / .55
	2010 (N= 593)	.04 (.19)	.68 (.07)	.10 (.31)	.44 (.06)	.19 (.40)	.37 (.07)	.47 (.06)	.54 / .50
	2012 (N= 615)	.04 (.21)	.70 (.09)	.13 (.34)	.52 (.06)	.30 (.46)	.51 (.07)	.55 (.06)	.56 / .51
	2014 (N= 666)	.04 (.19)	.62 (.07)	.14 (.34)	.40 (.06)	.20 (.40)	.42 (.06)	.45 (.05)	.51 / .49

Notes: MSA based on three dichotomous items for political participation. 'Item diff.' shows the frequency of each item with s.e. in parentheses. 'Item H' indicates the scalability coefficient for each item separately with s.e. in parentheses. 'Scale H' indicates the scalability coefficient for the final scale with s.e. in parentheses. Reliability indicated by 'rho/LCRC'. No violations of latent monotonicity and non-intersection found. ESS data 2002-2014.

an organization (see also Figure 2 below). This information provides additional evidence that the identity scale represents one and the same underlying concept (i.e., political participation) across all age groups and time points and thus forms a solid basis for meaningful comparisons of the equivalent scales to be built upon the identity set in the next step.

The second step of the 'identity-equivalence procedure' consists in adding further, age-group and time-point specific items to the identity scale. In this step, additional items are added as long as the properties of a Mokken Scale hold. More specifically, this implies that, in order to qualify as an extension of the identity scale, any of the four remaining items (i.e., wearing a badge or campaign sticker, signing a petition, taking part in a lawful demonstration, and boycotting products) has to meet the following criteria: It has to be positively correlated with the three constitutive items of the identity scale, exhibit a minimum scalability coefficient of 0.3 (item H), and lead to an overall degree of scalability of the resulting scale of at least 0.3 (scale H) (see also section 4). Any of the four items fulfilling these criteria is added to the identity scale to form equivalent scales of political participation that are comparable across age groups and time points. With this strategy, longer and more reliable scales of political participation can be reached that reflect the specific conditions of the respective age groups and time points while still being manifestations of the same underlying concept due to their inclusion of the same identity set. The results of this second step are summarized in Table 2.

In sixteen out of twenty-one cases the identity scales could be enriched with age-group or time-point specific items. For the adult age group, with the exception of 2008, the scale for political participation could be extended by three additional items (wearing a badge, signing a petition, demonstrating). The same holds true for the oldest age group in the years 2004, 2006, 2008 and 2012 as well as for the youngest age group in 2008. For the youngest age group it is further noteworthy that the identity scale could not be extended at all in the years 2004, 2010, 2012 and 2014. The item for boycotting products was not scalable in any of the age groups across time (item H <0.3). In thirteen out of twenty-one cases, the final equivalent scales of political participation establish weak scales (Scale H 0.3-0.39), while in the remaining eight cases they form medium scales (Scale H 0.4-0.49). More importantly, however, in all sixteen cases where additional items could be added, the reliability of the final equivalent scales in comparison to the identity scale could be improved.

Table 2 also provides information on the rank order of the individual modes of participation within the final equivalent scales of political participation. While this information is negligible for the construction of the equivalent scales itself, it provides some additional insights with regard to the differences in the participation profiles across age groups and time. As can be seen, even in those instances where the final equivalent scales are identical across the three age groups, the rank order of the individual modes differs between young, adult, and old people. Using the six-item equivalence scale as an example, we see that for young and adult people the least popular (or most 'difficult') mode of participation is working for a political party, whereas for the oldest age group it is wearing a badge. We also observe that contacting politicians is more difficult for young people as compared to adult and old people, while the opposite holds true for demonstrating. Yet, as the relative position of the three items of the identity set (which is the same across all respondents and years) does not change within the equivalence scales, these are still supposed to be comparable across age groups and time points.

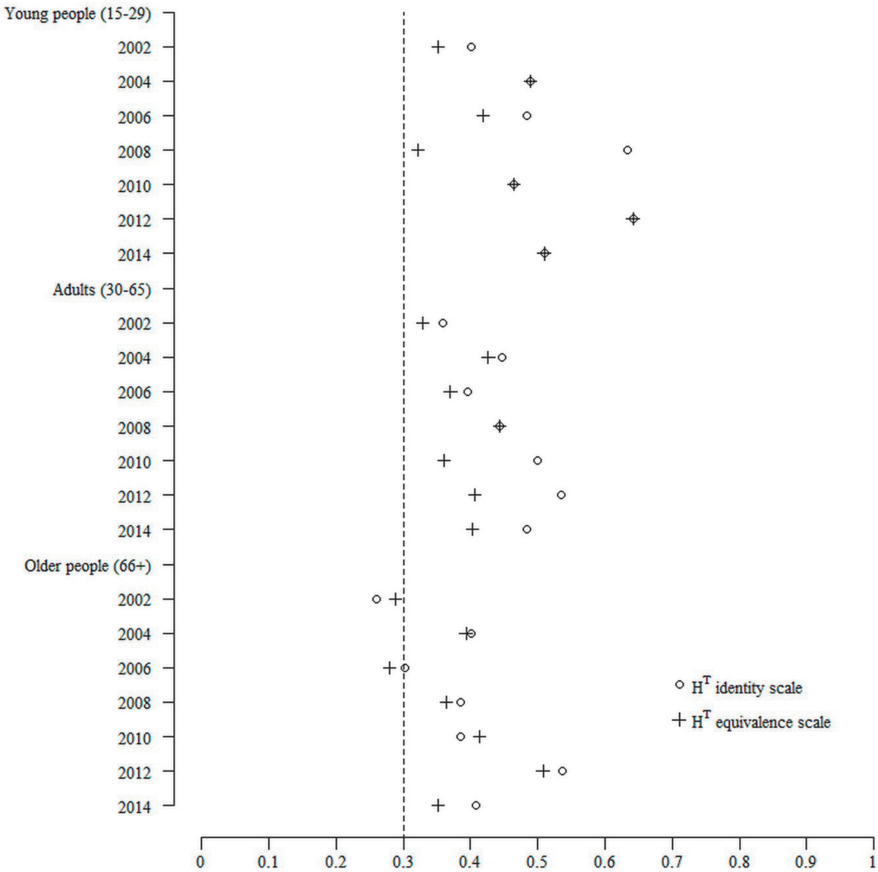
A more detailed investigation of the item ordering across age groups and time points is shown in Figure 2. Here we assessed whether the item rank orders as shown in Table 2 are the same for all respondents within a respective age group at a given point in time. In technical terms, we investigated the existence of an invari-

Table 2 Properties of the final equivalent scales of political participation across age groups and ESS waves

		Work for political party		Contacting politicians		Work for organisation		Wearing a badge		Signing a petition		Demonstrating			
Age group	Year	Item diff. (rank)	Item H	Item diff. (rank)	Item H	Item diff. (rank)	Item H	Item diff. (rank)	Item H	Item diff. (rank)	Item H	Item diff. (rank)	Item H	Scale H	Scale reliability
Young people (15-29)	2002	.03 (1)	.54 (.08)	.09 (3)	.30 (.06)	.17 (4)	.40 (.05)	.08 (2)	.35 (.06)	.30 (5)	.43 (.06)	--	--	.39 (.05)	.60 / .63
	2004	.03 (1)	.39 (.13)	.05 (2)	.35 (.09)	.18 (3)	.32 (.10)	--	--	--	--	--	--	.35 (.09)	.39 / .34
	2006	.01 (1)	.32 (.13)	.09 (2)	.34 (.06)	.17 (3)	.43 (.05)	--	--	.27 (4)	.37 (.06)	--	--	.38 (.05)	.55 / .71
	2008	.02 (1)	.54 (.10)	.09 (3)	.33 (.05)	.26 (5)	.32 (.05)	.06 (2)	.36 (.06)	.28 (6)	.36 (.05)	.10 (4)	.34 (.05)	.35 (.04)	.60 / .68
	2010	.03 (1)	.42 (.11)	.11 (2)	.36 (.07)	.25 (3)	.34 (.08)	--	--	--	--	--	--	.36 (.07)	.42 / .38
	2012	.04 (1)	.54 (.09)	.09 (2)	.37 (.08)	.32 (3)	.44 (.09)	--	--	--	--	--	--	.44 (.08)	.41 / .36
	2014	.03 (1)	.45 (.10)	.14 (2)	.36 (.07)	.31 (3)	.36 (.08)	--	--	--	--	--	--	.38 (.07)	.42 / .39
Adults (30-65)	2002	.04 (1)	.39 (.04)	.14 (4)	.33 (.02)	.19 (5)	.32 (.02)	.06 (2)	.38 (.03)	.32 (6)	.38 (.03)	.10 (3)	.31 (.03)	.34 (.02)	.61 / .66
	2004	.03 (1)	.46 (.03)	.13 (4)	.37 (.03)	.21 (5)	.40 (.03)	.04 (2)	.39 (.03)	.35 (6)	.46 (.03)	.07 (3)	.31 (.03)	.40 (.02)	.63 / .66
	2006	.04 (1)	.43 (.04)	.13 (4)	.39 (.03)	.21 (5)	.39 (.03)	.05 (2)	.40 (.04)	.30 (6)	.41 (.03)	.07 (3)	.35 (.03)	.40 (.02)	.63 / .69
	2008	.04 (1)	.58 (.06)	.18 (2)	.45 (.03)	.26 (3)	.44 (.03)	--	--	--	--	--	--	.47 (.03)	.56 / .54
	2010	.04 (1)	.42 (.04)	.17 (4)	.38 (.02)	.26 (5)	.35 (.02)	.05 (2)	.39 (.04)	.34 (6)	.33 (.03)	.09 (3)	.32 (.03)	.36 (.02)	.60 / .67
	2012	.05 (1)	.42 (.04)	.18 (4)	.37 (.03)	.32 (5)	.34 (.02)	.06 (2)	.32 (.04)	.39 (6)	.36 (.03)	.09 (3)	.33 (.03)	.36 (.02)	.59 / .64
	2014	.04 (1)	.39 (.04)	.16 (4)	.38 (.03)	.27 (5)	.34 (.02)	.05 (2)	.39 (.03)	.38 (6)	.36 (.03)	.09 (3)	.31 (.03)	.36 (.02)	.60 / .64

		Work for political party		Contacting politicians		Work for organisation		Wearing a badge		Signing a petition		Demonstrating			
Age group	Year	Item diff. (rank)	Item H	Item diff. (rank)	Item H	Item diff. (rank)	Item H	Item diff. (rank)	Item H	Item diff. (rank)	Item H	Item diff. (rank)	Item H	Scale H	Scale reliability
Older people (66+)	2002	.02 (2)	.47 (.07)	.08 (3)	.36 (.07)	.11 (4)	.34 (.08)	.02 (1)	.36 (.09)	--	--	--	--	.38 (.06)	.59 / .62
	2004	.02 (2)	.44 (.08)	.07 (4)	.46 (.07)	.14 (5)	.36 (.06)	.01 (1)	.33 (.13)	.19 (6)	.47 (.06)	.03 (3)	.35 (.08)	.41 (.06)	.63 / .69
	2006	.03 (2)	.48 (.07)	.08 (4)	.32 (.06)	.14 (5)	.37 (.05)	.02 (1)	.44 (.09)	.16 (6)	.37 (.06)	.03 (3)	.42 (.07)	.39 (.05)	.64 / .70
	2008	.03 (3)	.47 (.07)	.13 (4)	.40 (.05)	.18 (5)	.39 (.05)	.03 (1)	.43 (.07)	.19 (6)	.35 (.05)	.03 (2)	.45 (.08)	.40 (.04)	.66 / .74
	2010	.04 (2)	.54 (.07)	.10 (3)	.43 (.06)	.19 (4)	.38 (.07)	.03 (1)	.32 (.10)	--	--	--	--	.42 (.06)	.55 / .58
	2012	.04 (2)	.46 (.06)	.13 (4)	.43 (.06)	.30 (5)	.51 (.06)	.03 (1)	.50 (.08)	--	--	.05 (3)	.31 (.07)	.44 (.05)	.58 / .61
	2014	.04 (3)	.46 (.06)	.14 (4)	.32 (.05)	.20 (5)	.34 (.04)	.03 (1)	.41 (.07)	.21 (6)	.32 (.04)	.03 (2)	.40 (.06)	.36 (.04)	.61 / .65

Notes: MSA based on six dichotomous items for political participation. For the number of cases included in the analysis, see Table 1. 'Item diff.' shows the frequency of each item with its rank across all items in parentheses (1= most difficult/least popular). 'Item H' indicates the scalability coefficient for each item separately with s.e. in parentheses. 'Scale H' indicates the scalability coefficient for the final scale with s.e. in parentheses. Reliability indicated by 'rho/LCRC'. No violations of latent monotonicity and non-intersection found. ESS data 2002-2014.



Notes: For further information, see Ligtvoet et al. 2010; 2011.

Figure 2 Inspection of invariant item ordering (IIO) for political participation scales across three age groups and seven time points (H^T coefficients)

ant item ordering (IIO) across respondents as implied by the double monotonicity assumption in MSA (cf. Ligtvoet et al., 2010, 2011; Sijtsma, Meijer, & van der Ark, 2011; Watson et al., 2014). The existence of IIO implies that the item response functions of any pair of items do not intersect and are sufficiently different from each other to speak of a meaningful order of items across respondents. For the inspection of IIO, Ligtvoet et al. (2010, 2011) have proposed the coefficient H^t which should exhibit a minimum value of 0.3 in order to draw meaningful conclusions about the existence of IIO of the items within a Mokken Scale (see also Sijtsma, Meijer, & van der Ark, 2011; Watson et al., 2014, 74-5).

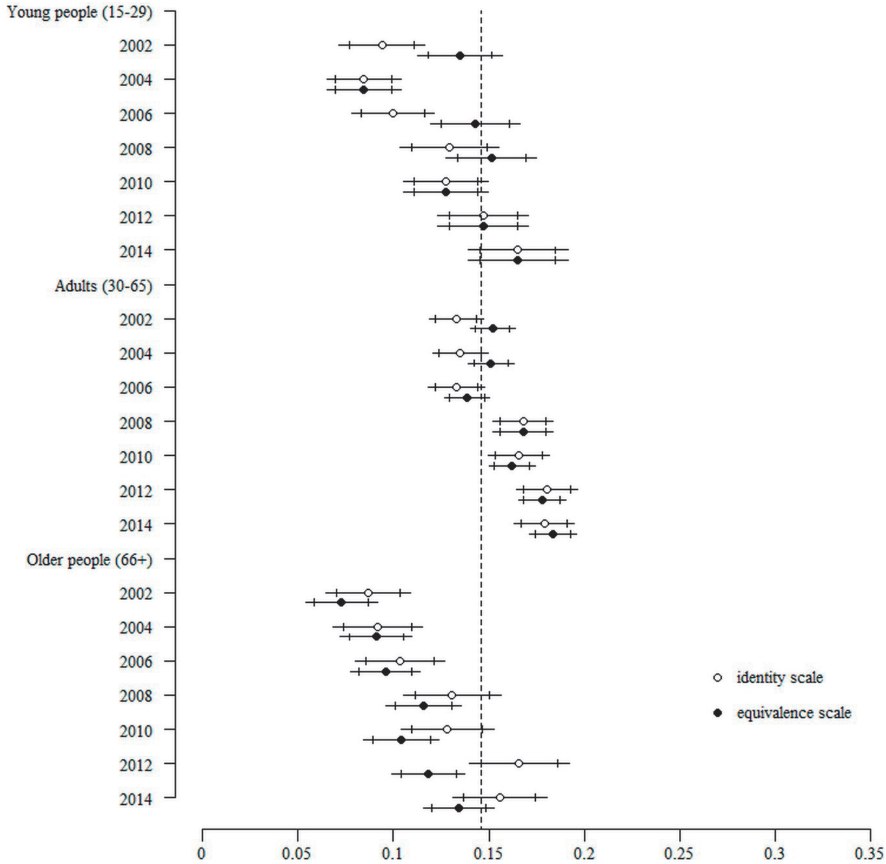
Figure 2 plots the respective H^t coefficients for the identity and equivalence scales of political participation across age groups and time points. With only two exceptions, all coefficients exceed the critical value of 0.3 so that we can speak of an invariant item ordering and a meaningful order of the different modes of political participation across all respondents within the same age group at a given point in time. Both exceptions refer to the oldest age group (2002 and 2006) where the item response functions for the different items are too close to each other (H^t<0.3) to convey any meaningful message about the order of the items across all respondents.

As a final summary of the 'identity-equivalence procedure' for political participation, Figure 3 provides a descriptive overview of the final identity and equivalence scales across age groups and time points. To account for the varying number of items in the final equivalence scales across age groups and time points, we have standardized all scales to range from 0-1 (cf. van Deth, 1986, 269).

Three observations seem to be notable. First, it can be seen that the identity and equivalence scales yield varying participation levels. Especially for the youngest age group the equivalence scales reveal higher participation levels than the identity scales (see 2002, 2006 and 2008), while for the oldest age group the opposite can be observed. For the group of adults, the equivalence scales sometimes result in higher and sometimes in lower participation levels than the identity scales. Second, there appears to be an increasing trend in the levels of political participation over time for young, adult, and old people alike. Accordingly, for all three age groups, the average participation levels are higher in 2014 than 2002. Third, comparing the levels of political participation across young, adult, and old people, the oldest age group clearly is the least politically active. However, more interesting from the perspective of youth participation research is the observation that young and adult people in fact show quite similar levels of political participation. With the exception of 2004, young people's political participation does not deviate significantly from the average participation levels of the overall population. Using equivalent instruments of political participation that are based on a common identity set thus provides us with a less gloomy picture about young people's political participation than relying on the commonly employed strategy of applying identical instruments.

Summary and Discussion

In applying the 'identity-equivalence procedure' for political participation across different age groups and time points, this study offers a re-assessment of young people's political participation in Germany. As Cammaerts et al. have pointed out, "much of the existing social science literature, as well as many journalistic accounts of the supposedly low turnout of young people in elections, assumes that



Notes: ESS data for the years 2002-2014, data weighted using post-stratification weights. The vertical line shows the average level of political participation across all three age groups and time points as measured by the equivalence scale.

Figure 3 Average levels of political participation across three age groups and seven time points (means with 99% and 95% CIs)

young people today are simply fed up with politics per se and not interested in the political questions facing their communities or their countries. However, much of this literature fails to provide convincing empirical evidence for such claims and critiques" (2014, 650). In this study, we argue that, in order to arrive at meaningful conclusions about young people's political participation, its specificities in comparison with adult people, as well as its developments over time, 'construct-equivalent' rather than identical instruments of political participation across different age groups and over time should be used.

What are the main insights of the 'identity-equivalence procedure' for political participation across age groups and time points? First, the (empirical) structure of the concept of political participation does not reflect commonly employed types of political participation, such as the distinction between institutionalized and non-institutionalized participation. For all age groups and time points under investigation, MSA yields a single, uni-dimensional scale of political participation. In light of this finding, the commonly employed strategy of many previous studies of simply applying well-known distinctions between different types of political participation without checking their empirical suitability is at least questionable. Second, while the 'identity-equivalence procedure' shows a generally uni-dimensional structure of political participation, the concrete composition of the final equivalent scales of political participation as well as the rank order of the different participation modes within these scales vary across age groups and over time. Overall, the equivalence scales contain more items for adult people and are more stable in their composition over time when compared to young and old people. This finding might indeed be a reflection of a more narrow conception of politics held by young people as pointed out in previous research (cf. Quintelier, 2007, 177; O'Toole et al., 2003, 52). In any case, it shows that the concrete manifestation of the concept political participation differs across age groups and time. Simply applying identical (rather than equivalent) instruments of political participation for young and adult people as well as different time points thus appears to be an ill-suited strategy to arrive at meaningful conclusions about the levels and trends of political participation. Third, regarding the levels and trends of political participation, the results for our final equivalent scales show an increase in participation levels over time that is observable for all age groups. These results are clearly at odds with the conventional wisdom stating that young people are less politically active than adults and are becoming more and more politically apathetic and disengaged as time passes by (cf. Henn & Foard, 2014, 361). Judging from the results based on our equivalent scales of political participation, the future prospects of (German) democracy are not as shady as suggested in some previous studies of youth political participation.

What are the implications of the 'identity-equivalence procedure' for political participation across age groups and time points? In light of the results presented, a central question concerns the analytical value of commonly employed conceptions and types of political participation, such as the distinction between conventional and unconventional or institutionalized and non-institutionalized participation. As indicated earlier, for none of our age groups and time points under consideration the 'identity-equivalence procedure' as implemented by MSA yields a solution that consists of two (or more) scales and that could be indicative of any of the types of political participation mentioned above. Does this mean that we can completely eschew these commonly employed conceptions of political participation? Such a conclusion would certainly be premature. First, it is clear that (cross-national) surveys such as the ESS are limited in the number and the variety of items to be included in the survey. Constructing time-series data for a stable set of items logically comes at the expense of including new items into a survey when interviewing time is limited. This establishes a possible problem, as surveys such as the ESS are limited in their capability to adapt to recent changes and developments concerning political participation. As a consequence, survey items for newer modes of participation, such as 'guerilla gardening' or 'reclaim-the-street parties' (cf. van Deth, 2014), which might form the basis of a second dimension of political participation, are not available in the ESS. Hence, it might be possible that the uni-dimensionality of our equivalent scales establishes an artefact of the particular items used in the present analysis.[4] While there is certainly no easy answer to this problem, cross-national surveys such as the ESS sooner or later have to find a way to adapt to and cover changes in the empirical realities of concepts such as political participation. Second, the uni-dimensionality found for our equivalent participation scales might also be a direct consequence of the underlying logic of the 'identity-equivalence procedure'. As the procedure requires a common identity set that represents a valid scale across all subgroups considered, it might have obscured other, more-dimensional structures of political participation. However, since our goal was to establish 'construct-equivalent' scales of political participation for young and adult people over time, we did not inspect any scales that were not based on a common identity set for all age groups and time points.

What are the implications of the findings for comparative survey research in general and participation research in particular? Researchers investigating differences and similarities in the political behavior of young and adult people over time should ensure that (1) they use reliable samples including *both young and adult people,* (2) they track both groups *over time,* and (3) the *measurement of political participation* is equivalent across age groups as well as over time. Questions of measurement equivalence in the area of comparative survey research usually arise in the context of establishing equivalent instruments *across countries* (cf. Przeworksi & Teune, 1966; van Deth, 1986; Garcia Albacete, 2011; Linssen et al., 2014).

4 To investigate this argument, we checked the robustness of the results presented in Table 2 by repeating the same analysis with a broader set of participation items that is only available for the first wave of the ESS in 2002. The additional four items encompass (1) buycotting products, (2) donating money to a political organization, (3) participating in an illegal protest, and (4) taking part in a referendum. The results confirm the uni-dimensional scale of political participation across all age groups. For young and old people, the robustness check even yields the exact same equivalence scales as shown in Table 2. For adult people, the equivalence scale can be extended by the items for buycotting and donating money. Detailed results of the robustness check are available upon request.

As this study has pointed out, similar considerations concerning the equivalence of instruments may also apply if the main objective is to draw meaningful conclusions about differences and similarities *between different societal groups and points in time*. Accordingly, future studies on political participation and beyond should be (more) attentive to the fact that the analysis of one and the same phenomenon may require the usage of equivalent rather than identical instruments.

References

Bakker, T.P., & de Vreese, C.H. (2011). Good News for the Future? Young People, Internet Use, and Political Participation. *Communication Research,* 38, 451-70. Doi: https://doi.org/10.1177/0093650210381738

Cammaerts, B., Bruter, M., Banaji, S., Harrison, S., & Anstead, N. (2014). The Myth of Youth Apathy: Young Europeans Critical Attitudes Toward Democratic Life. *American Behavioral Scientist,* 58, 645-64. Doi: https://doi.org/10.1177/0002764213515992

Gaiser, W., de Rijke, J., & Spannring, R. (2010). Youth and Political Participation – Empirical Results for Germany within a European Context. *Young,* 18, 427-50. Doi: https://doi.org/10.1177/110330881001800404

Garcia Albacete, G. (2011). *Continuity or Generational Change? A Longitudinal Study of Young People's Political Participation in Western Europe*. Dissertation manuscript, University of Mannheim.

Hao, X., Wen, N., & George, C. (2014). News Consumption and Political and Civic Engagement among Young People. *Journal of Youth Studies,* 17, 1221-38. Doi: http://dx.doi.org/10.1080/13676261.2014.901490

Henn, M., & Foard, N. (2014). Young People, Political Participation and Trust in Britain. *Parliamentary Affairs,* 65, 47-67. Doi: https://doi.org/10.1093/pa/gsr046

Ligtvoet, R., van der Ark, L.A., Te Marvelde, J.M., & Sijtsma, K. (2010). Investigating an Invariant Item Ordering for Polytomously Scored Items. *Educational and Psychological Measurement,* 70, 578-95. Doi: https://doi.org/10.1177/0013164409355697

Ligtvoet, R., van der Ark, L.A., Bergsma, W.P., & Sijtsma, K. (2011). Polytomous Latent Scales for the Investigation of the Ordering of Items. *Psychometrika,* 76, 200-16.

Linssen, R., Schmeets, H., Scheepers, P., & te Grotenhuis, M. (2014). Trends in Conventional and Unconventional Political Participation in Europe, 1981-2008. In C. Eder, I.C. Mochmann & M. Quandt (Eds), *Political Trust and Disenchantment with Politics* (pp. 31-58). Leiden: Brill.

Marien, S., Hooghe, M., & Quintelier, E. (2010). Inequalities in Non-institutionalised Forms of Political Participation: A Multi-Level Analysis of 25 Countries. *Political Studies,* 58, 187-213. Doi: 10.1111/j.1467-9248.2009.00801.x

Martin; A. (2012). Political Participation among the Young in Australia: Testing Dalton's Good Citizen Thesis. *Australian Journal of Political Science,* 47, 211-26. Doi: http://dx.doi.org/10.1080/10361146.2012.677003

Milbrath, L.W. (1965). *Political Participation*. Chicago: Rand McNally.

Mokken, R. J. (1971). *A Theory and Procedure of Scale Analysis.* The Hague: Mouton & Co.

Mycock, A., & Tonge, J. (2012). The Party Politics of Youth Citizenship and Democratic Engagement. *Parliamentary Affairs,* 65, 138-61. Doi: https://doi.org/10.1093/pa/gsr053

Neufeind, M., Jiranek, P., & Wehner, T. (2014). Beyond Skills and Structure: Justice Dispositions as Antecedents of Young Citizens' Volunteering and Political Participation. *Journal of Community and Applied Social Psychology,* 24, 278-95. Doi: 10.1002/casp.2166

O'Toole, T., Lister, M., Marsh, D., Jones, S., & McDonagh, A. (2003). Tuning Out or Left Out? Participation and Non-Participation among Young People. *Contemporary Politics,* 9, 45-61. Doi: http://dx.doi.org/10.1080/1356977032000072477

Przeworski, A., & Teune, H. (1966). Equivalence in Cross-National Research. *Public Opinion Quarterly,* 30, 551-68.

Quintelier, E. (2007). Differences in Political Participation between Young and Old People. *Contemporary Politics,* 13, 165-80. Doi: http://dx.doi.org/10.1080/13569770701562658

Quintelier, E. (2008). Who is Politically Active: The Athlete, the Scout Member or the Environmental Activist? *Acta Sociologica,* 51, 355-70. Doi: https://doi.org/10.1177/0001699308097378

Schnaudt, C., Walter, S. and Popa, S.A. (2016). 'Subnational and National Territorial Identification'. In Westle, B. and Segatti, P. (eds), *European Identity in the Context of National Identity: Questions of Identity in Sixteen European Countries in the Wake of the Financial Crisis*. Oxford: Oxford University Press, 63-92.

Schnaudt, C., Weinhardt, M., Fitzgerald, R., & Liebig, S. (2014). The European Social Survey: Contents, Design and Research Potential. *Schmollers Jahrbuch,* 134, 487-506. Doi: https://doi.org/10.3790/schm.134.4.487

Schneekloth, U. (2015). Jugend und Politik: Zwischen positivem Gesellschaftsbild und anhaltender Politikverdrossenheit. In Shell Deutschland Holding (Ed.), *Jugend 2015. Eine pragmatische Generation im Aufbruch* (pp. 153-200). Frankfurt am Main: Fischer.

Sijtsma, K., & Molenaar, I.W. (2002). *Introduction to Nonparametric Item Response Theory.* Thousand Oaks, CA: Sage.

Sijtsma, K., Meijer, R.R., & van der Ark, L.A. (2011). Mokken Scale Analysis as Time Goes By: An Update for Scaling Practitioners. *Personality and Individual Differences,* 50, 31-37. Doi: https://doi.org/10.1016/j.paid.2010.08.016

Sloam, J. (2013). "Voice and Equality": Young People's Politics in the European Union. *West European Politics,* 36, 836-58. Doi: http://dx.doi.org/10.1080/01402382.2012.749652

Sloam, J. (2014). New Voice, Less Equal: The Civic and Political Engagement of Young People in the United States and Europe. *Comparative Political Studies,* 47, 663-88. Doi: https://doi.org/10.1177/0010414012453441

Soler-i-Marti, R., & Ferrer-Fons, M. (2015). Youth Participation in Context: The Impact of Youth Transition Regimes on Political Action Strategies in Europe. *Sociological Review,* 63(S2), 92-117. Doi: 10.1111/1467-954X.12264

Stolle, D., Hooghe, M., & Micheletti, M. (2005). Politics in the Supermarket: Political Consumerism as a Form of Political Participation. *International Political Science Review,* 26, 245-69. Doi: 10.1177/0192512105053784

van der Ark, L. A. (2007). Mokken Scale Analysis in R. *Journal of Statistical Software,* 20, 1-19. Doi: http://dx.doi.org/10.18637/jss.v020.i11

van der Ark, L.A. (2012). New Developments in Mokken Scale Analysis in R. *Journal of Statistical Software,* 48, 1-24. Doi: http://dx.doi.org/10.18637/jss.v048.i05

van der Ark, L.A., van der Palm, D.W., & Sijtsma, K. (2011). A latent class approach to estimating test-score reliability. *Applied Psychological Measurement, 35*, 380-392. Doi: https://doi.org/10.1177/0146621610392911 van Deth, J.W. (1986). A Note on Measuring Political Participation in Comparative Research. *Quality & Quantity,* 120, 261-72. Doi: https://doi.org/10.1007/BF00227430

van Deth, J.W. (2014). A Conceptual Map of Political Participation. *Acta Politica,* 49, 349-67. Doi: https://doi.org/10.1057/ap.2014.6

van Schuur, W.H. (2003). Mokken Scale Analysis: Between the Guttman Scale and Parametric Item Response Theory. *Political Analysis,* 11, 139-63. Doi: 10.1093/pan/mpg002

Verba, S., & Nie, N.H. (1972). *Participation in America. Political Democracy and Social Equality*. New York: Harper & Row.

Verba, S., Schlozman, K.L., & Brady, H.E. (1995). *Voice and Equality. Civic Voluntarism in American Politics*. Cambridge: Harvard University Press.

Vissers, S., & Stolle, D. (2014). The Internet and New Modes of Political Participation: Online versus Offline Participation. *Information, Communication & Society,* 17, 937-55. Doi: http://dx.doi.org/10.1080/1369118X.2013.867356

Watson, R., Wang, W., Thompson, D.R., & Meijer, R.R. (2014). Investigating Invariant Item Ordering in the Mental Health Inventory: An Illustration of the Use of Different Methods. *Personality and Individual Differences,* 66, 74-78. Doi: https://doi.org/10.1016/j.paid.2014.03.017

Wray-Lake, L., & Hart, D. (2012). Growing Social Inequalities in Youth Civic Engagement? Evidence from the National Election Study. *PS: Political Science & Politics,* 45, 456-61. Doi: https://doi.org/10.1017/S1049096512000339

Information for Authors

Methods, data, analyses (mda) publishes research on all questions important to quantitative methods, with a special emphasis on survey methodology. In spite of this focus we welcome contributions on other methodological aspects.

Manuscripts that have already been published elsewhere or are simultaneously submitted to other journals will not be considered. As a rule we do not restrict authors' rights. All rights remain with the author, and articles in mda are published under the CC-BY open-access license.

Mda aims for a quick peer-review process. All papers submitted to mda will first be screened by the editors for general suitability and then double-blindly reviewed by at least two reviewers. The decision on publication is made by the editors based on the reviews. The editorial team will contact the authors by email with the result at the latest eight weeks after submission; if the reviews have not been received by then, we provide a status update with a new target date.

When preparing a paper for submission, please consider the following guidelines:

- Please submit your manuscript via www.mda.gesis.org.
- The total length of the manuscript shall not exceed 10.000 words.
- Manuscripts should…
    - be written in English, using American English spelling. Please use correct grammar and punctuation. Non-native English speakers should consider a professional language editing prior to publication.
    - be typed in a 12 pt Roman font, double-spaced throughout.
    - be submitted as MS Word documents.
    - start with a cover page containing the title of the paper and contact details / affiliations of the authors, but be anonymized for review otherwise.
    - should be anonymized ("blinded") for review.
- Please also send us an abstract of your paper (approx. 300 words), a front page with a brief biographical note (no longer than 250 words as supplementary file), and a list of 5-7 keywords for your paper.
- Acceptable formats for Graphics are
    - pdf
    - jpg (uncompressed, high quality)
- Please ensure a resolution of at least 300 dpi and take care to send hiqh-quality graphics. Line art images should have a resolution of 500-1000 dpi. Please note that we cannot print color images.
- The type area of our journal is 11.5 cm (width) x 18.5 cm (height). Please consider this when producing tables or graphics.
- Footnotes should be used sparingly.
- Please number the headings of your article. Doing so will make the work of the layout editor easier.

- If your text includes formulas we would like to ask you to upload your text also as a PDF, additional to the Word document.
- By submitting a paper to mda the authors agree to make data and program routines available for purposes of replication.
- Response rates in research papers or research notes, where population surveys are analyzed, should be calculated according to AAPOR standard definitions.
- Please follow the APA guideline when structuring and formating your work.

When preparing in-text references and the list of references please also follow the APA guidelines:

Entire Book:

Groves, R. M., & Couper, M. P. (1998). *Nonresponse in household interview surveys*. New York: John Wiley & Sons.

Journal Article (with DOI):

Klimoski, R., & Palmer, S. (1993). The ADA and the hiring process in organizations. *Consulting Psychology Journal: Practice and Research*, 45(2), 10-36. doi:10.1037/1061-4087.45.2.10

Journal Article (without DOI):

Abraham, K. G., Helms, S., & Presser, S. (2009). How social processes distort measurement: The impact of survey nonresponse on estimates of volunteer work in the United States. *American Journal of Sociology*, 114(4), 1129-1165.

Chapter in an Edited Book:

Dixon, J., & Tucker, C. (2010). Survey nonresponse. In P. V. Marsden & J. D. Wright (Eds.), *Handbook of Survey Research*. Second Edition (pp. 593-630). Bingley: Emerald.

Internet Source (without DOI):

Lewis, O., & Redish, L. (2011). *Native American tribes of Wisconsin*. Retrieved April 19, 2012, from the Native Languages of the Americas website: www.native-languages.org/wisconsin.htm

For more information, please consult the Publication Manual of the American Psychological Association (Sixth ed.).

gesis
Leibniz Institute for the Social Sciences

ISSN 1864-6956 (Print)
ISSN 2190-4936 (Online)